UNIVERSITAT POLITECNICA DE CATALUNYA

The effect of noise and sample size in the performance of an unsupervised feature relevant determination method for manifold learning

by

Jorge Sebastian Velazco Brao

A thesis submitted in partial fulfillment for the
degree of Master in Artificial Intelligence

in the

Facultat d'Informatica de Barcelona
Departament de Llenguatges i Sistemes Informatics

June 2008

"I never teach my pupils; I only attempt to provide the conditions in which they can learn."

Albert Einstein

UNIVERSITAT POLITECNICA DE CATALUNYA

Abstract

Facultat d'Informatica de Barcelona
Departament de Llenguatges i Sistemes Informatics

Master in Artificial Intelligence

by Jorge Sebastian Velazco Brao

The research on unsupervised feature selection is scarce in comparison to that for supervised models, despite the fact that this is an important issue for many clustering problems. An unsupervised feature selection method for general Finite Mixture Models was recently proposed and subsequently extended to Generative Topographic Mapping (GTM), a manifold learning constrained mixture model that provides data clustering and visualization. Some of the results of previous research on this unsupervised feature selection method for GTM suggested that its performance may be affected by insufficient sample size and by noisy data. In this thesis, we test in detail such limitations of the method and outline some techniques that could provide an at least partial solution to the negative effect of the presence of uninformative noise. In particular, we provide a detailed account of a variational Bayesian formulation of feature relevance determination for GTM.

Acknowledgements

I would like to express my gratitude to all those who gave me the possibility to complete this thesis. To my thesis supervisor Dr. Alfredo Vellido whose help, stimulating suggestions and encouragement helped me throughout my research.

My colleagues from the Master and PhD. program of Artificial Intelligence that supported me in my research work. I want to thank them all for their help, support, interest and valuable suggestions.

Jorge Velazco is a master student and acknowledges funding from the Spanish Ministry of Education and Science (MEC) "Becas para cursar estudios de Másteres Oficiales en el curso 2007-2008".

Contents

List of Figures

Abbreviations

GTM	**G**enerative **T**opographic **M**apping
FS	**F**eature **S**election
FRD	**F**eature **R**elevance **D**etermination
ML	**M**aximum **L**ikelihood
EM	**E**xpectation **M**aximization
SML	**S**tatistical **M**achine **L**earning
FMM	**F**inite **M**ixture **M**odels
ARD	**A**utomatic **R**elevance **D**etermination
SOM	**S**elf **O**rganizing **M**ap
SRT	**S**ingle **R**egularization **T**erm
SMS	**S**elective **M**apping **S**moothing

Dedicated to my mother...

Chapter 1

Introduction

1.1 Introduction

The fields of machine learning and statistics coexist with data analysis as a common target and they overlap in what has come to be defined as Statistical Machine Learning. An example of this can be found in Finite Mixture Models (FMM), which are flexible and robust methods for multivariate data clustering [1]. The addition of visualization capabilities would benefit these models in many application scenarios, helping to provide intuitive cues about data structural patterns. One way to endow FMM with data visualization is by constraining the mixture components to be centred in a low-dimensional manifold embedded into the multivariate data space, as in Generative Topographic Mapping (GTM) [2]. This is a manifold learning model for simultaneous data clustering and visualization.

The interpretability of the clustering results provided by GTM becomes difficult when the analyzed data sets consist of a large number of features. This limitation can be overcome with methods to estimate the ranking of the data features according to their relative relevance, leading to feature selection (FS). The research on unsupervised FS is scarce in comparison to that for supervised models, despite the fact that FS becomes an issue of paramount importance for many clustering problems, regardless the unavailability of class labels. The interpretability of the clusters obtained by unsupervised methods would be improved by their description in terms of a reduced subset of relevant variables.

An important advance on unsupervised FS for FMM was presented in [3] and recently extended to GTM in [4] and to one of its variants for time series analysis in [5]. This method was preliminarily assessed in [6], where some of the results suggested that the performance of the method may be degraded by characteristics of the data such as

insufficient sample size and the presence of noise. In this thesis, we provide evidence of the limitations of the method through controlled experiments using mostly synthetic but also some real data.

1.2 Motivation and objectives

The method for Feature Relevance Determination using GTM (FRD-GTM) described in [7] was preliminarily and partially assessed in [8], where some of the results suggested that its performance may be to some extent degraded by characteristics of the data such as insufficient sample size and the presence of uninformative noise. In this thesis, we provide evidence of the limitations of the method through a battery of experiments using mostly synthetically generated data, which allows us to control the nature of the data in terms of expected relevance for clustering.

In its basic formulation, the GTM is trained within the Maximum Likelihood (ML) framework using Expectation-Maximization (EM), permitting the occurrence of data overfitting unless regularization is included, a major drawback when modelling noisy data. This limitation indeed extends to FRD-GTM. Statistical Machine Learning (SML) provides a unified principled framework for machine learning methods and helps to overcome some of their limitations, such as data overfitting due to the presence of noise.

In the last chapter of this thesis, we outline the basics of some possible methods to deal with the presence of noise in the analyzed datasets. GTM, as a SML method, allows the formulation of principled extensions, such as those providing active model regularization. Some regularization methods for GTM described in [9, 10] are based on Bayesian evidence approaches. They could be extended to FRD-GTM. Alternatively, a variational Bayesian approach of the GTM was recently introduced in [11, 12] to endow the model with regularization capabilities based on variational techniques, showing very promising results. In this thesis, we provide the basic formulation of a FRD method for Variational GTM. We also describe the potential use of a variation of GTM based on the use of geodesic distances.

In summary, the objectives for this thesis are:

1. An exhaustive assessment of the effects of insufficient sample size and the presence of uninformative noise in the performance of unsupervised FRD using GTM.
2. The description of some potential alternatives to address the negative effects of the presence of uninformative noise, including the detailed formulation of a variational Bayesian method for FRD-GTM.

Chapter 2

Basic Background Theory

2.1 Feature Relevance and Feature Selection

Feature Selection (FS) is the straightest of the strategies for dimensionality reduction, consisting in the selection of a subset of inputs, discarding the remainder. This approach can be useful if there are inputs which carry little relevant information for the solution of the problem at hand, or if, alternatively, there are very strong correlations between sets of inputs.

Any procedure for feature selection must be based on at least two components. First, a criterion must be defined by which it is possible to gauge whether the relevance of a subset of features is better than the relevance of another. Second, a systematic procedure must be found for searching through candidate subsets of features. Some of the benefits of feature selection include:

- Facilitating data visualisation and understanding as part of multivariate, high-dimensional data exploration.
- Reducing measurement efforts and information storage requirements.
- Reducing computational load.
- Defying the curse of dimensionality to improve prediction performance.

The problems of FRD and the subsequent FS based on it can be studied in the context of supervised learning. In such setting, which has been thoroughly studied, a data feature is said to be relevant (and it is eventually selected) only if its absence (or its absence in combination with the absence of others) worsens significantly the classification or

predictive performance of the defined model. It is beyond the purpose of this chapter to provide a complete review of the many approaches and techniques of supervised FS. Such reviews can be found elsewhere [13]. Suffice it to say that the two main available approaches are the wrapper and filter techniques.

In short, **wrapper** feature selection consists in building a classifier with an aim to achieve the highest predictive accuracy possible and select the features used by the classifier as the optimal features, it weigh up subsets of features according to their usefulness to a given predictor. There is another model known, as the **filter** model, which is base on distance and information measures, selecting features by ranking them with correlation coefficients.

FS and FRD for unsupervised learning, even if sharing the dimensionality reduction objective of their supervised counterparts, are far less investigated problems. Here, the relevance is not longer related to a label or target variable, maybe because this label is not available at all or only partially available, or even because the labels are available but we are interested in the exploration of the structure of the data themselves.

Various unsupervised feature ranking criteria can be considered, including, but not limited to, saliency, entropy, smoothness, density and reliability [13]. One reason, even if not the only, to consider a feature as salient is if it has a high variance or a large range, as compared to others. A feature has high entropy if the distribution of examples it generates is uniform and, therefore, irrelevant for the definition of informative structure. A feature is in a high-density region if it is highly correlated with many other features. Finally, a feature is reliable if the measurement error bars computed by repeating measurements are small, as compared to the inherent variability of the feature values.

2.2 Finite Mixture Models

2.2.1 Unsupervised clustering by learning mixtures of Gaussians

Finite Mixture Models are Statistical Machine Learning (SML) methods for multivariate data clustering. SML provides a unified principled framework for machine learning methods and helps to overcome some of their limitations. Bayesian probability theory, in particular, has important modeling implications. For instance, it requires modeling assumptions, including parameters and prior distributions, to be made explicit, avoiding arbitrary modelling decisions; it also automatically satisfies the likelihood principle and provides a natural framework to handle uncertainty.

In mixture models, the observed data are assumed to be samples of a combination or finite mixture of $k = 1, \ldots, K$ components or underlying distributions, weighted by unknown priors $P(k)$. Given a D-dimensional dataset $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$, consisting of N random observations, the corresponding mixture density is defined as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(x|k;\Theta_k)\, P(k)\,, \tag{2.1}$$

where each mixture component k is parameterized by Θ_k. For continuous data, the choice of Gaussian distributions is a rather straightforward option due to their computational convenience [14], in which case

$$p(\mathbf{x}|k;\mu_k,\Sigma_k) = (2\pi)^{-D/2}\,|\Sigma_k|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right\}, \tag{2.2}$$

where the adaptive parameters Θ_k are the mean vector and the covariance matrix of the D-variate distribution for each mixture component, namely μ_k and Σ_k. Their Maximum Likelihood estimates can be obtained using the EM algorithm and, for that, first we define the complete log-likelihood as

$$L_c(\mu,\Sigma|\mathbf{X}) = \log\prod_{n=1}^{N} p(\mathbf{x}_n) = \sum_{n=1}^{N}\log\sum_{k=1}^{K} p(\mathbf{x}_n|k;\mu_k,\Sigma_k)\,P(k)\,. \tag{2.3}$$

In the context of the EM algorithm, we can introduce the binary indicator variables $\mathbf{Z} = \{\mathbf{z}_k\}_{k=1}^{K}$, with $\mathbf{Z}_k = (z_{k1},\ldots,z_{kN})$, which reflect our ignorance of which mixture component k is responsible for the generation of data observation n. The complete log-likelihood can now be expressed as

$$L_c\left(\mu,\Sigma|\mathbf{X},\mathbf{Z}\right)=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{kn}\log\left[p\left(\mathbf{x}_n|k;\mu_k,\Sigma_k\right)P\left(k\right)\right]. \tag{2.4}$$

The indicators $\mathbf{Z}$ are effectively treated as missing data and, following the iterative EM procedure, the re-estimation of the adaptive parameters μ_k, Σ_k requires the maximization of the expected log-likelihood $E[L_c\left(\mu,\Sigma|\mathbf{X},\mathbf{Z}\right)|\mathbf{X},\mu_k,\Sigma_k]$.

The expectation of each of the indicators in $\mathbf{Z}$, which is the probability of a mixture component k being responsible for data observation $\mathbf{x}_n$ (also known as responsibility r_{kn}) can be written as:

$$r_{kn}=p\left(k|\mathbf{x}_n,\mu_k,\Sigma_k\right)=\frac{\left|\Sigma_k\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\mathbf{x}_n-\mu_k\right)^T\Sigma_k^{-1}\left(\mathbf{x}_n-\mu_k\right)\right\}P\left(k\right)}{\Sigma_{k'=1}^{K}\left|\Sigma_{k'}\right|^{-1/2}\exp\left\{-\frac{1}{2}\left(\mathbf{x}_n-\mu_{k'}\right)^T\Sigma_{k'}^{-1}\left(\mathbf{x}_n-\mu_{k'}\right)\right\}P\left(k'\right)} \tag{2.5}$$

With this, in the maximization step, the update formulae for μ_k, Σ_k are obtained as:

$$\hat{\mu}_k=\frac{\Sigma_{n=1}^{N}r_{kn}\mathbf{x}_n}{\Sigma_{n=1}^{N}r_{kn}} \tag{2.6}$$

$$\hat{\Sigma}_k=\frac{\Sigma_{n=1}^{N}r_{kn}\left(\mathbf{x}_n-\hat{\mu}_k\right)\left(\mathbf{x}_n-\hat{\mu}_k\right)^T}{\Sigma_{n=1}^{N}r_{kn}} \tag{2.7}$$

2.2.2 Feature Relevance Determination in Gaussian Mixture Models

The problem of feature relative relevance determination for GMM was recently addressed in [3]. Feature relevance in this unsupervised setting is understood as the likelihood of a feature being useful to define the data clustering structure. In that sense, it becomes a soft version of a FS method: no feature is actually meant to be discarded because none is likely to be either completely useful or useless. However, the resulting relevance ranking can be the basis of an *a posteriori* selection. A similar counterpart procedure for supervised models is Automatic Relevance Determination (ARD: [15, 16]).

Formally, the saliency of feature d is defined as $\rho_d=P\left(\eta_d=1\right)$, where $\eta=\left(\eta_1,\ldots,\eta_D\right)$ is a further set of binary indicators that, like $\mathbf{Z}$, can be integrated in the EM algorithm as missing variables. A value of $\eta_d=1$ indicates the full relevance of feature d. According to this definition, the mixture density in Eq. 2.1 can be rewritten as:

$$p(\mathbf{x}) = \sum_{k=1}^{K} P(k) \prod_{d=1}^{D} \{\rho_d p(x_d|k;\Theta_k) + (1-\rho_d)\, q(x_d|\lambda_d)\} \tag{2.8}$$

Notice that this entails the assumption that features are conditionally independent given a mixture component, which is equivalent to the assumption of a diagonal covariance matrix. The distribution p would be a univariate version of Eq. 2.2, and the relevance of feature d would be given by ρ_d; consequently, a feature d would be considered as irrelevant, with *irrelevance* $(1-\rho_d)$, if, for all mixture components, $p(x_d|k;\Theta_{kd}) = q(x_d|\lambda_d)$, where $q(x_d|\lambda_d)$ is a common density followed by feature d, or common mixture component. Notice that this is tantamount to say that the distribution for feature d does not follow the cluster structure defined by the GMM. This common component should reflect any prior knowledge we might have regarding irrelevant features, or otherwise take the form of a general, uninformative distribution.

The maximum likelihood criterion can now be made explicit as the estimation of those model parameters that maximize the complete log-likelihood

$$L_c = \sum_{n=1}^{N} \log \sum_{k=1}^{K} P(k) \prod_{d=1}^{D} (\rho_d p(x_d|k;\Theta_k) + (1-\rho_d)\, q(x_d|\lambda_d)) \tag{2.9}$$

which can be accomplished using the EM algorithm (For details, see [3]). The probability of a component k being the generator of observation n : r_{kn} , is computed in the expectation step of the algorithm as:

$$r_{kn} = \frac{P(k)\,\Pi_d \{\rho_d p(x_{nd}|k;\Theta_{kd}) + (1-\rho_d)\, q(x_{nd}|\lambda_d)\}}{\Sigma_{k'} P(k')\,\Pi_d \{\rho_d p(x_{nd}|k';\Theta_{k'd}) + (1-\rho_d)\, q(x_{nd}|\lambda_d)\}}. \tag{2.10}$$

Then, the maximization step provides update expressions for the components' priors $P(k) \equiv \alpha_k$, for the means and variances associated to each feature d in $p(\cdot|\cdot)$ and $q(\cdot|\cdot)$, as well as for the relevance parameter ρ_d:

$$\hat{\alpha}_k = \Sigma_n r_{kn}/N \tag{2.11}$$

$$\hat{\mu}_{\Theta_{kd}} = \frac{\Sigma_n \frac{\rho_d P(x_{nd}|k;\Theta_{kd})}{\rho_d P(x_{nd}|k;\Theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn} x_{nd}}{\Sigma_n \frac{\rho_d P(x_{nd}|k;\Theta_{kd})}{\rho_d P(x_{nd}|k;\Theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn}} \tag{2.12}$$

$$\hat{\Sigma}_{\Theta_{kd}} = \frac{\Sigma_n \frac{\rho_d P(x_{nd}|k;\Theta_{kd})}{\rho_d P(x_{nd}|k;\Theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn} \left(x_{nd} - \hat{\mu}_{\Theta_{kd}}\right)^2}{\Sigma_n \frac{\rho_d P(x_{nd}|k;\Theta_{kd})}{\rho_d P(x_{nd}|k;\Theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn}} \tag{2.13}$$

$$\hat{\mu}_{\lambda_d} = \frac{\Sigma_n \Sigma_k \left(\frac{(1-\rho_d) q(x_{nd}|\lambda_d)}{\rho_d P(x_{nd}|k;\Theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} \right) x_{nd}}{\Sigma_n \Sigma_k \frac{(1-\rho_d) q(x_{nd}|\lambda_d)}{\rho_d P(x_{nd}|k;\Theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn}} \tag{2.14}$$

$$\hat{\Sigma}_{\lambda_d} = \frac{\Sigma_n \Sigma_k \left(\frac{(1-\rho_d) q(x_{nd}|\lambda_d)}{\rho_d P(x_{nd}|k;\Theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} \right) (x_{nd} - \hat{\mu}_{\lambda_d})^2}{\Sigma_n \Sigma_k \frac{(1-\rho_d) q(x_{nd}|\lambda_d)}{\rho_d P(x_{nd}|k;\Theta_{kd}) + (1-\rho_d) q(x_{nd}|\lambda_d)} r_{kn}} \tag{2.15}$$

$$\hat{\rho}_d = \frac{1}{N} \Sigma_{n,k} \frac{\rho_d P\left(x_{nd}|k;\Theta_{kd}\right)}{\rho_d P\left(x_{nd}|k;\Theta_{kd}\right) + \left(1-\rho_d\right) q\left(x_{nd}|\lambda_d\right)} r_{kn} \tag{2.16}$$

Chapter 3

Unsupervised Feature Relevance Determination for GTM

The Finite Mixture Models described in the previous chapter have settled in recent years as a standard for statistical modelling. Gaussian Mixture Models, in particular, have received especial attention for their computational convenience [14] to deal with multivariate continuous data. The usefulness of these models is reinforced by the wide spectrum of their applications.

In practice, GMM suffer from several shortcomings that may limit their applicability. One of them is their lack of multivariate data visualization capabilities. Data visualization can be especially important in the exploratory data analysis. The GTM model was originally conceived as a constrained GMM that circumvected this limitation by enabling the visualization of multivariate data on a low dimensional space. In this chapter, we provide the basic theoretical definition of GTM and its extension to perform feature relevance determination: the FRD-GTM.

3.1 GTM: The Generative Topographic Mapping

The GTM [2] was originally formulated both as a probabilistic alternative to Kohonen's SOM [17] and as a constrained mixture of distributions. It is precisely its constrained definition that allows overcoming the data and cluster visualization limitations of general finite mixture models. The GTM is a non-linear latent variable model that defines a mapping from a low dimensional latent space onto the multivariate data space. The mapping is carried through by a set of basis functions generating a (mixture) density

distribution. The functional form of this mapping is defined as a generalized linear regression model:

$$y_d(\mathbf{u}, \mathbf{W}) = \sum_m^M \phi_m(\mathbf{u}) w_{md}, \tag{3.1}$$

where ϕ is a set of M basis functions $\phi(\mathbf{u}) = (\phi_1(\mathbf{u}), \ldots, \phi_M(\mathbf{u}))$ that were originally defined as spherically symmetric Gaussians $\phi_m(\mathbf{u}) = \exp\left\{-\frac{\|\mathbf{u}-\mu_m\|^2}{2\sigma^2}\right\}$, with μ_m the centres of the Gaussians and σ their common width; $\mathbf{W}$ is the matrix of adaptive weights w_{md} that defines the mapping; and $\mathbf{u}$ is a point in latent space. In order to achieve computational tractability and to provide an alternative to the clustering and visualization space defined by the characteristic SOM lattice, the latent space of the GTM is discretized as a regular grid of K latent points $\mathbf{u}_k$ defined by the probability

$$P(\mathbf{u}) = 1/K \sum_{k=1}^{K} \delta(\mathbf{u} - \mathbf{u}_k), \tag{3.2}$$

where δ is the Kronecker's delta. The probability distribution for a data point $\mathbf{x}$, induced by the latent distribution in Eq. 2.9, takes the form of isotropic Gaussian noise and, given the adaptive parameters of the model, which are the matrix $\mathbf{W}$ and the inverse variance of the Gaussians β, it can be written as:

$$p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta) = \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2 \|\mathbf{x} - \mathbf{y}\|^2\right\}, \tag{3.3}$$

where the elements of $\mathbf{y}$ are given by Eq. 3.1. Marginalizing over the latent points and using Eq. 3.2, we obtain

$$p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta) = \int p(\mathbf{x}|\mathbf{u}, \mathbf{W}, \beta) P(\mathbf{u}) d\mathbf{u} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2 \|\mathbf{x} - \mathbf{y}_k\|^2\right\} \tag{3.4}$$

According to this general description, the GTM is a constrained mixture of Gaussians in the sense that all the components of the mixture are equally weighted by the term $1/K$, all components share a common variance β^{-1} (therefore $\sum = \beta^{-1}\mathbf{I}$), and the centres of the Gaussian components $\mathbf{y}_k = \phi(\mathbf{u}_k)\mathbf{W}$ do not move independently from each other, as they are limited by the mapping definition to lie on a low dimensional manifold embedded in the D-dimensional space. Notice that, given the common variance constrain, the GTM

complies by definition with the assumption that features are conditionally independent given a mixture component, expressed in section 2.2.2.

The complete log-likelihood can now be defined as:

$$L_c\left(\mathbf{W},\beta|\mathbf{X}\right)=\sum_{n=1}^{N}\log\left\{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\beta}{2\pi}\right)^{D/2}\exp\left\{-\beta/2\left\|\mathbf{x}_n-\mathbf{y}_k\right\|^2\right\}\right\} \tag{3.5}$$

As for GMM, we can resort to the EM algorithm to obtain the Maximum Likelihood estimates of the adaptive parameters $\mathbf{W}$ and β. Defining once again as $\mathbf{Z}$ the indicators describing our lack of knowledge of which latent point $\mathbf{u}_k$ is responsible for the generation of data point $\mathbf{x}_n$, the complete log-likelihood can be rewritten as

$$L_c\left(\mathbf{W},\beta|\mathbf{X},\mathbf{Z}\right)=\sum_{n=1}^{N}\sum_{k=1}^{K}z_{kn}\log\left[\left(\frac{\beta}{2\pi}\right)^{D/2}\exp\left\{-\beta/2\left\|\mathbf{x}_n-\mathbf{y}_k\right\|^2\right\}\right] \tag{3.6}$$

The expected value of z_{kn} is now an special case of Eq. 2.5

$$r_{kn}=P\left(k|\mathbf{x}_n,\mathbf{W},\beta\right)=\frac{\exp\left\{-\frac{\beta}{2}\left\|\mathbf{x}_n-\mathbf{y}_k\right\|^2\right\}}{\sum_{k'=1}^{K}\exp\left\{-\frac{\beta}{2}\left\|\mathbf{x}_n-\mathbf{y}_{k'}\right\|^2\right\}} \tag{3.7}$$

The update expressions for $\mathbf{W}$ and β are computed in the maximization step. We obtain $\mathbf{W}^{new}$ as the solution of the following system of equations in matricial form:

$$\mathbf{\Phi}^T\mathbf{G}\mathbf{\Phi}\mathbf{W}^{new}-\mathbf{\Phi}^T\mathbf{R}\mathbf{X}=0, \tag{3.8}$$

where $\mathbf{\Phi}$ is a $K\times M$ matrix with elements $\phi_{km}=\phi_m\left(\mathbf{u}_k\right)$; $\mathbf{R}$ is the responsibility matrix, with elements r_{kn}; and $\mathbf{G}$ is a matrix with values

$$g_{kk'}=\left\{\begin{array}{cc}\sum_{n=1}^{N}r_{kn} & k=k'\\ 0 & k\neq k'\end{array}\right\}$$

.

Notice that Eq. 3.8 is equivalent to Eq. 2.6, given that the component centres for the GTM are described by $\mathbf{Y}=\mathbf{\Phi}\mathbf{W}$.

The update expression for β is:

$$(\beta^{new})^{-1} = \frac{1}{ND}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{kn} \left\|\mathbf{x}_n - \mathbf{y}_k\right\|^2 \tag{3.9}$$

See [2] for further details on these calculations.

3.2 Feature Relevance Determination for GTM: The FRD-GTM

The approach to FRD in unsupervised models described in section 2.2.2 can be transferred to the standard Gaussian GTM. It has to be born in mind that, to some extent, the relevance of a feature depends on the number of clusters defined by a given solution. Considering the GTM strictly from its definition as a constrained mixture model, each of the points of the latent space sampling defined by Eq. 3.2 can be thought as the generator of a single data cluster. For data visualization purposes, the number of latent points is left rather unconstrained in the usual GTM definition. Therefore, the FRD method applied to GTM should be understood as a constrained one in as far as it is meant to reach a compromise between its own ability as detector of feature relevance in clustering structure, and the data visualization capabilities of the GTM. In other words, for FRD-GTM, individual features are relevant in the sense that they explain the specific clustering structure provided by GTM, and not necessarily the unconstrained clustering structure of the data.

For FRD-GTM, the complete log-likelihood in Eq. 3.5 becomes:

$$L_c\left(\mathbf{W}, \mathbf{w}_0, \beta, \beta_0, \mathbf{p}|\mathbf{X}\right) = \sum_{n=1}^{N} \log\left\{\frac{1}{K}\sum_{k=1}^{K}\prod_{d=1}^{D}\left(a_{knd} + b_{knd}\right)\right\}, \tag{3.10}$$

where

$$a_{knd} = \rho_d \left(\beta/2\pi\right)^{1/2} \exp\left(-\frac{\beta}{2}\left(x_{nd} - \Sigma_m \phi_m\left(\mathbf{u}_k\right) w_{md}\right)^2\right), \tag{3.11}$$

$$b_{knd} = \left(1-\rho_d\right) \left(\beta_{0,d}/2\pi\right)^{1/2} \exp\left(-\frac{\beta_{0,d}}{2}\left(x_{nd} - \phi_0\left(\mathbf{u}_0\right) w_0\right)^2\right), \tag{3.12}$$

and $\beta_0 \equiv \{\beta_{0,1}, \ldots, \beta_{0,D}\}$;$\rho_0 \equiv \{\rho_1, \ldots, \rho_D\}$. The common component requires the definition of two extra adaptive parameters $\mathbf{w}_0$ and β_0, so that $\mathbf{y}_0 = \phi_0\left(\mathbf{u}_0\right) w_0$.

This common component accounts for data observations that the constrained mixture components cannot explain well; in other words, data observations that do not fit with the cluster structure described by these components. This approach is not unlike the one commonly used to deal with the presence of atypical data observations (outliers) when fitting Gaussian mixtures, which entails the inclusion of an additional component with a uniform distribution. This can be circumvented by the fitting of Student t-distribution mixtures [18], which has also been done for GTM [19]. The FRD method presented in this report, though, differs from the former on its featurewise approach.

Resorting again to the EM algorithm, we rewrite the complete log-likelihood of the model as:

$$L_c\left(\mathbf{W}, \mathbf{w}_0, \beta, \beta_0, \mathbf{p}|\mathbf{X}, \mathbf{Z}\right) = \Sigma_{n,k} r_{kn} \sum_{d=1}^{D} \log\left(a_{knd} + b_{knd}\right) \tag{3.13}$$

where the expected *responsibiblity* in Eq. 3.7 becomes:

$$r_{kn} = p\left(k|\mathbf{x}_n, \mathbf{W}, \mathbf{w}_0, \beta, \beta_0, \rho\right) = \frac{\prod_{d=1}^{D}\left(a_{knd} + b_{knd}\right)}{\sum_{k'=1}^{K}\prod_{d=1}^{D}\left(a_{k'nd} + b_{k'nd}\right)}. \tag{3.14}$$

The maximization of the expected log-likelihood for GTM yields the following update formulae for parameters ρ, $\mathbf{W}$, β, $\mathbf{w}_0$ and β_0:

$$\rho_d^{new} = \frac{1}{N}\Sigma_{n,k} r_{kn} u_{knd}, \tag{3.15}$$

where

$$u_{knd} = \frac{a_{knd}}{a_{knd} + b_{knd}}. \tag{3.16}$$

$$\beta^{new} = \frac{\Sigma_{n,k} r_{kn} \Sigma_d u_{knd}}{\Sigma_{n,k} r_{kn} \Sigma_d u_{knd}\left(x_{nd} - \Sigma_m \phi_m\left(\mathbf{u}_k\right) w_{md}\right)^2} \tag{3.17}$$

$$\beta_{0,d}^{new} = \frac{\Sigma_{n,k} r_{kn} v_{knd}}{\Sigma_{n,k} r_{kn} v_{knd}\left(x_{nd} - \phi_0\left(\mathbf{u}_0\right) w_{0,d}\right)^2}, \tag{3.18}$$

where

$$v_{knd} = \frac{b_{knd}}{a_{knd} + b_{knd}}. \tag{3.19}$$

For fully relevant ($\rho_d \to 1$) features, the common component variance vanishes:

$(\beta_{0,d})^{-1} \to 0$. We now obtain, for each feature d, the elements of matrix $\mathbf{W}^{new}$ as the solution of the following system of equations in matricial form:

$$\mathbf{\Phi}^T \mathbf{G}^* \mathbf{\Phi} \mathbf{W}_d^{new} - \mathbf{\Phi}^T \mathbf{R}^* \mathbf{X}_d = 0, \tag{3.20}$$

where $\mathbf{R}^*$ has elements $r^*_{kn} = u_{knd} * r_{kn}$ for a given feature d^* with r_{kn} given by Eq. 3.14, and $\mathbf{G}^*$ has elements

$$g^*_{kk'} = \left\{ \begin{array}{cc} \sum_{n=1}^{N} r^*_{kn} & k = k' \\ 0 & k \neq k' \end{array} \right\}$$

.

Notice the similarity of Eq. 3.20 and Eq. 3.8. Similarly, we obtain $\mathbf{w}_0^{new}$, featurewise, as the solution of:

$$\Phi^T g^* \phi_0 \mathbf{w}_{0,d}^{new} - \phi^T \mathbf{r}^* \mathbf{X}_d = 0, \tag{3.21}$$

where $\mathbf{r}^*$ has elements $r^* = \Sigma_k r^*_{kn} = \Sigma_k v_{knd} * r_{kn}$ for a given feature d^*, and $g^* = \Sigma_{n,k} r^*_{kn}$.

Note that the expression $u_{knd} r_{kn}$ could be considered as the *responsibility* of the constrained mixture component k for generating feature d of a data observation n. Correspondingly, expression $v_{knd} r_{kn}$ could actually be considered as the *lack of responsibility* of the constrained mixture component k for generating feature d of a data observation n.

Chapter 4

Experiments

4.1 Introduction

The main objective of this thesis consists in the investigation of the possible effects of noise and sample size in the performance of unsupervised feature selection using mixture models. The study of such effects is relevant for many reasons:

- It allows reducing the dimensionality of the data, redefining the datasets through a smaller subset of features.
- It facilitates multivariate data visualisation and an easier understanding of the knowledge extracted by the model.
- It has the potential to reduce measurement and computational storage requirements.
- It has the potential to reduce development and deployment times.

4.2 Experimental Settings

The results of statistically principled models for probability density estimation, such as GTM and its variants, are bound to be affected, in one way or another, by sample size and by the presence of uninformative noise in the data. Here, we assess such effects on the FRD-GTM model described in the previous chapter. For that, data with very specific characteristics are required. We mostly use synthetic sets similar to those in [3] for comparative purposes.

4.2.1 Data Description

4.2.1.1 Trunk data (SYNTH1)

The first synthetic set (hereafter referred to as *SYNTH1*) is a variation on the *Trunk* data set used in [3], and was designed for its 10 features to be in decreasing order of relevance. It consists of data sampled from two Gaussians $N(\mu_1, \mathbf{I})$ and $N(\mu_2, \mathbf{I})$, where: $\left(\mu_1 = 1, \frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{2d-1}}, \ldots, \frac{1}{\sqrt{19}}\right)$ and $\mu_1 = -\mu_2$. Samples of *SYNTH1* of different sizes, from 100 to 10,000 points, were used in this study to test the effect of sample size. In order to test the effect of noise, four increasing levels of Gaussian noise, of standard deviations 0.1, 0.2, 0.5, and 1, were added to the 10 original features of *SYNTH1*, for a given sample size.

4.2.1.2 Experiment 1 (SYNTH2)

The second dataset (hereafter referred to as *SYNTH2*) consists of a contrasting combination of features: the first two define four neatly separated Gaussian clusters with centres located at $(0, 3)$, $(1, 9)$, $(6, 4)$ and $(7, 10)$; they are meant to be relatively relevant. The next four features are Gaussian noise and, therefore, rather irrelevant in terms of defining cluster structure. Similar experiments to the ones devised for *SYNTH1* were designed for this dataset.

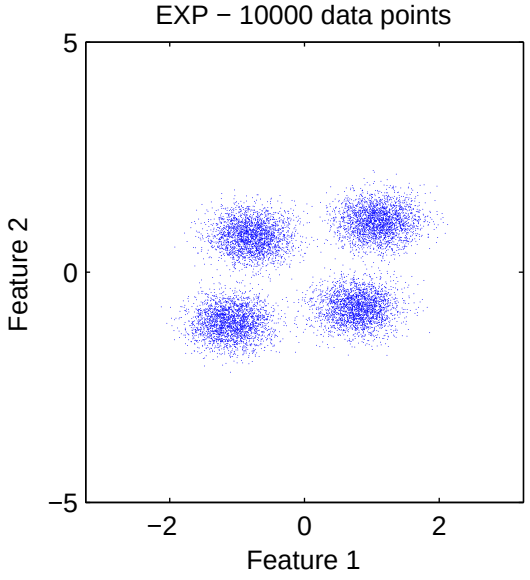

FIGURE 4.1: Graphical representation of the two first informative features of an illustrative sample of 10,000 points from Experiment 1 (SYNTH2)

4.2.1.3 SYNTH3. 4 Gaussians (close centres)

Consists of data points from a mixture of four equiprobable Gaussians $N(m_i, I)$ and $i = \{1, 2, 3, 4\}$, where $m_1 = (1\ 4)$, $m_2 = (1\ 8)$, $m_3 = (5\ 4)$ and $m_4 = (5\ 8)$. Four "noisy" features (sampled from a $N(0, 1)$ distribution) are appended to these data.

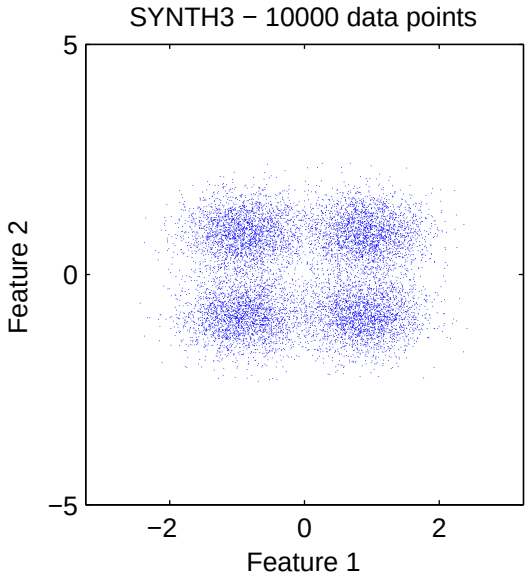

FIGURE 4.2: Graphical representation of the two first informative features of an illustrative sample of 10,000 points from SYNTH3

4.2.1.4 SYNTH4. 4 Gaussians (flat)

Consists of data points from a mixture of four equiprobable Gaussians $N(m_i, I)$ and $i = \{1, 2, 3, 4\}$, where $m_1 = (0\ 3)$, $m_2 = (0\ 9)$, $m_3 = (6\ 3)$, $m_4 = (6\ 9)$ and

$$cov = \begin{vmatrix} 1 & 1.15 \\ 1.15 & 3 \end{vmatrix}$$

Four "noisy" features (sampled from a $N(0, 1)$ distribution) are appended to these data.

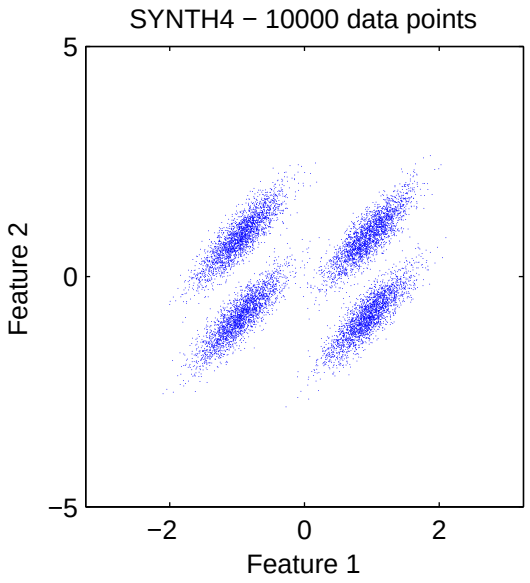

FIGURE 4.3: Graphical representation of the two first informative features of an illustrative sample of 10,000 points from SYNTH4

4.2.1.5 SYNTH5. 4 Gaussians (lineal)

Consists of data points from a mixture of four equiprobable Gaussians $N(m_i, I)$ and $i = \{1, 2, 3, 4\}$, where $m_1 = (0\ 3)$, $m_2 = (6\ 3)$, $m_3 = (12\ 3)$ and $m_4 = (18\ 3)$. Four "noisy" features (sampled from a $N(0, 1)$ distribution) are appended to these data.

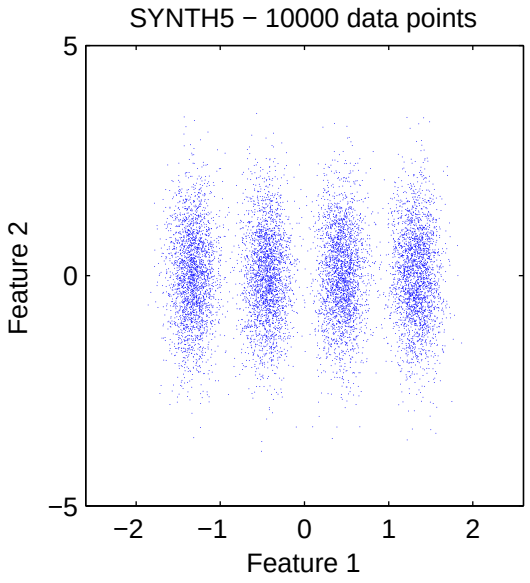

FIGURE 4.4: Graphical representation of the two first informative features of an illustrative sample of 10,000 points from SYNTH5

4.2.1.6 SYNTH6. 6 Gaussians

Consists of data points from a mixture of six equiprobable Gaussians $N(m_i, I)$ and $i = \{1, 2, 3, 4\}$, where $m_1 = (0\ 3)$, $m_2 = (1\ 9)$, $m_3 = (6\ 4)$, $m_4 = (7\ 10)$, $m_5 = (12\ 5)$ and $m_6 = (13\ 11)$. Four "noisy" features (sampled from a $N(0, 1)$ distribution) are appended to these data.

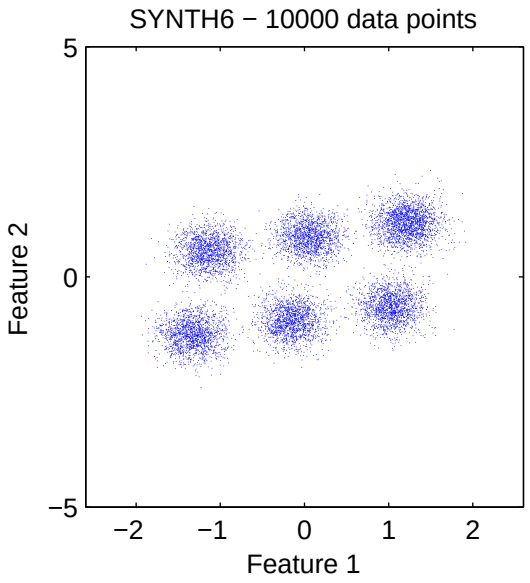

FIGURE 4.5: Graphical representation of the two first informative features of an illustrative sample of 10,000 points from SYNTH6

4.2.1.7 SYNTH7. 8 Gaussians

Consists of data points from a mixture of eight equiprobable Gaussians $N(m_i, I)$ and $i = \{1, 2, 3, 4\}$, where $m_1 = (0\ 3)$, $m_2 = (1\ 9)$, $m_3 = (6\ 3)$, $m_4 = (7\ 9)$, $m_5 = (12\ 3)$, $m_6 = (13\ 9)$, $m_7 = (18\ 3)$ and $m_8 = (19\ 9)$. Four "noisy" features (sampled from a $N(0, 1)$ distribution) are appended to these data.

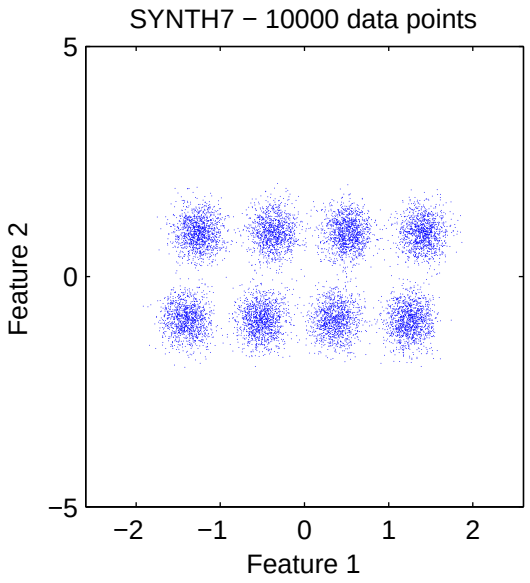

FIGURE 4.6: Graphical representation of the two first informative features of an illustrative sample of 10,000 points from SYNTH7

4.2.1.8 SYNTH8. 6 Gaussians (lineal)

Consists of data points from a mixture of six equiprobable Gaussians $N(m_i, I)$ and $i = \{1, 2, 3, 4\}$, where $m_1 = (0\ 3)$, $m_2 = (1\ 9)$, $m_3 = (6\ 3)$, $m_4 = (7\ 9)$, $m_5 = (12\ 3)$ and $m_6 = (13\ 9)$. Four "noisy" features (sampled from a $N(0, 1)$ distribution) are appended to this data.

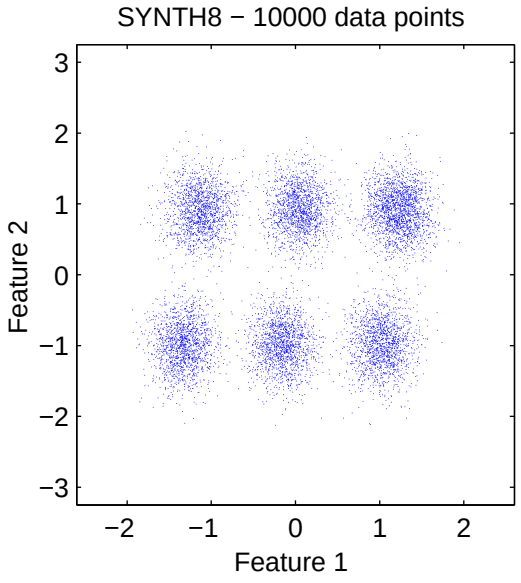

FIGURE 4.7: Graphical representation of the two first informative features of an illustrative sample of 10,000 points from SYNTH8

4.2.1.9 SYNTH9. 8 Gaussians (diagonal)

Consists of data points from a mixture of eight equiprobable Gaussians $N(m_i, I)$ and $i = \{1, 2, 3, 4\}$, where $m_1 = (0\ 3)$, $m_2 = (1\ 9)$, $m_3 = (6\ 4)$, $m_4 = (7\ 10)$, $m_5 = (12\ 5)$, $m_6 = (13\ 11)$, $m_7 = (18\ 6)$ and $m_8 = (19\ 12)$. Four "noisy" features (sampled from a $N(0, 1)$ distribution) are appended to these data.

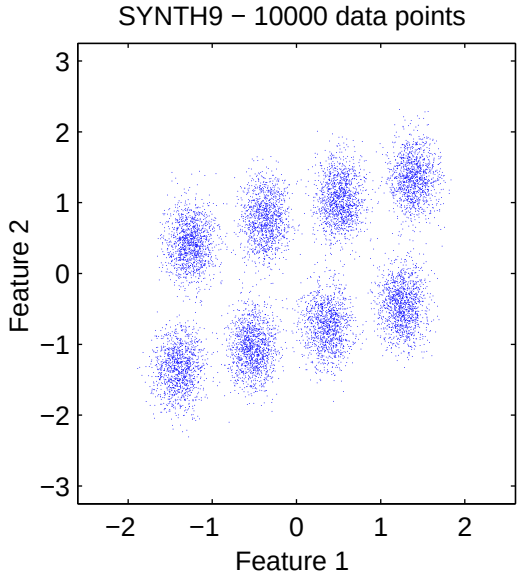

FIGURE 4.8: Graphical representation of the two first informative features of an illustrative sample of 10,000 points from SYNTH9

4.2.1.10 Real Data: the *Ionosphere* Dataset

The well known *Ionosphere* data set from the UCI Machine Learning Repository will also be used for analysis. It contains radar data consisting of 351 instances and 34 features, the latter consisting of 17 pairs of values. Each pair is formed by the real and complex parts of the values of an autocorrelation function for a pulse number of the radar system signal. The first pair was removed due to uninformative character of its complex part. The ionosphere data were originally meant for classification, as they can be ascribed to one of two categories or classes: "bad radar returns" and "good radar returns". Such classes, in turn, indicate the lack of or the existence of ionosphere structure.

4.2.2 FRD-GTM settings

One of the modelling decisions to be made when setting up a GTM model in general, and FRD-GTM in particular, is that of the model architecture. This takes the form

of the choice of discretization of the latent visualization space, described in the previous chapter. As happens with its close relative, the Self-Organizing Map (SOM: [17]) method, the grid of GTM latent centres can take different layouts and sizes. Previous research [6, 7] has shown the reasonable lack of sentivity of FRD-GTM to changes in the model's architecture. In this thesis, we investigated different architectures without finding any significant differences in performance regarding the results that concern the experimental hypotheses.

Therefore, and for the sake of brevity, we report in detail here only the results corresponding to a fixed grid of GTM latent centres with a square layout of 3×3 nodes (i.e., 9 constrained mixture components). The corresponding grid of basis functions was fixed to a 2×2 layout. The FRD-GTM parameters $\mathbf{W}$ and $\mathbf{w}_0$ were initialized with small random values sampled from a normal distribution. Saliencies were initialized at $\rho_d = 0.5, \forall d, d = 1, \ldots, D$. The EM algorithm for parameter optimization by log-likelihood maximization is prone to lead the algorithm towards local minima and initializing the algorithm with random weights in multiple runs is meant to at least alleviate this shortcoming of the optimization method, making the comparative experiments more reliable.

For the experiments with the *Ionosphere* data, the grid of GTM latent centres was fixed to square layouts of 5×5 and 10×10 nodes (i.e., 25 or 100 constrained mixture components). The corresponding grid of basis functions ϕ_m was fixed to 3×3 and 5×5 layouts.

4.2.3 Experimental hypotheses

This chapter aims to assess the effect of noise and sample size on the performance of the FRD-GTM model. Such goal opens a large breadth of possible experimental designs that is unreasonable to implement in full. Selecting a finite number of synthetic datasets has already narrowed the choice. Then, a number of sample sizes from 10,000 down to 100 data points was considered as a sensible selection to illustrate the effect of sample size. Not all those that were used are reported in this chapter, but the selection suffices for the assessment. We first hypothesize (H1) that the feature relevance ranking estimated by FRD-GTM for all these datasets will deteriorate gradually as sample size decreases.

Then again, different types of noise might have been considered for study, but, in this thesis, we focus on Gaussian noise (noise that has the p.d.f. of the Gaussian or normal distribution). Two different approaches were used to gauge the effect of noise. Firstly, for all experiments, uninformative Gaussian noise of different and increasing standard deviations was added to the informative data features. Although a wider array of values

was used, here we report results for added noise of standard deviations 0.1 (herein referred to as Level 1), 0.2 (Level 2), 0.5 (Level 3), and 1 (Level 4). Secondly, for the experiments concerning different combinations of Gaussian distributions, new features, consisting of just Gaussian noise and therefore uninformative, randomly sampled from $N(0,1)$, were added to the original dataset features in different numbers. Only the results for 3 and 6 new added features are later reported in the appendices.

According to these settings it is also hypothesized that the feature relevance ranking will deteriorate in proportion to the level of noise added to the original data features (H2.1) and that the feature relevance ranking will deteriorate in proportion to the number of uninformative noisy features added to the original data features (H2.2).

4.3 Experimental results and discussion

We first report and discuss the results of the assessment of the effect of sample size on the unsupervised saliency estimation by FRD-GTM. This is followed by the report and discussion of the results of the assessment of the effect of noise on the same estimation. All the reported experiments correspond to the settings detailed in section 4.2.

4.3.1 The effect of sample size on the unsupervised saliency estimation by FRD-GTM

The FRD ranking results for the Trunk data (SYNTH1) are shown in Figures 4.9 and 4.10, for sample sizes from 10,000 down to 100 points. A deterioration of the results is clearly observed for datasets of less than 1,000 points. This deterioration takes two forms: Firstly, a breach of the expected monotonic decrease of the mean feature saliencies. Secondly, a neat increase of uncertainty in the results, illustrated in Figures 4.9 and 4.10 in the form of bigger bars of the standard deviation of the estimated saliencies. As a result, the confidence on the validity of the results for small sample sizes decreases considerably. According to these results, the hypothesis *H1*, outliend in section 4.2, is at least partially supported.

The FRD ranking results for SYNTH2, again for sample sizes from 10.000 down to 100 points, are shown in Figures 4.11 and 4.12. The problem of four Gaussians, quite well separated and arranged in a quite symmetric layout, is a much easier easier problem for the FRD-GTM model, and this is reflected by the fact that the saliency estimated for the two first features is higher than that estimated for the rest of the features, even for a sample size as small as 100 points. A deterioration of the saliency estimation is nevertheless evident for the smallest of the sample sizes investigated. This is consistent with the results for SYNTH1 and, again, *H1* is partially supported.

To avoid cluttering the text with an excessive amount of figures, those corresponding to the following experiments are relegated to Appendix A. The FRD ranking results for SYNTH3, for the same sample sizes, are shown in Figures A.1 and A.2. This should be a harder problem for the model than the one posed by SYNTH2, given that the four artificially generated normally distributed clusters have centres that are much closer to each other than those of SYNTH2 and, therefore, their level of overlapping is higher. As a result, you might expect the relevance of the first two features to be more difficult to assess. This is the case, and it is reflected by the fact that the saliency estimated for the two first features is overall lower than that estimated for SYNTH2 . Despite the fact that the saliency of the first two features is differentially higher than the saliency of the

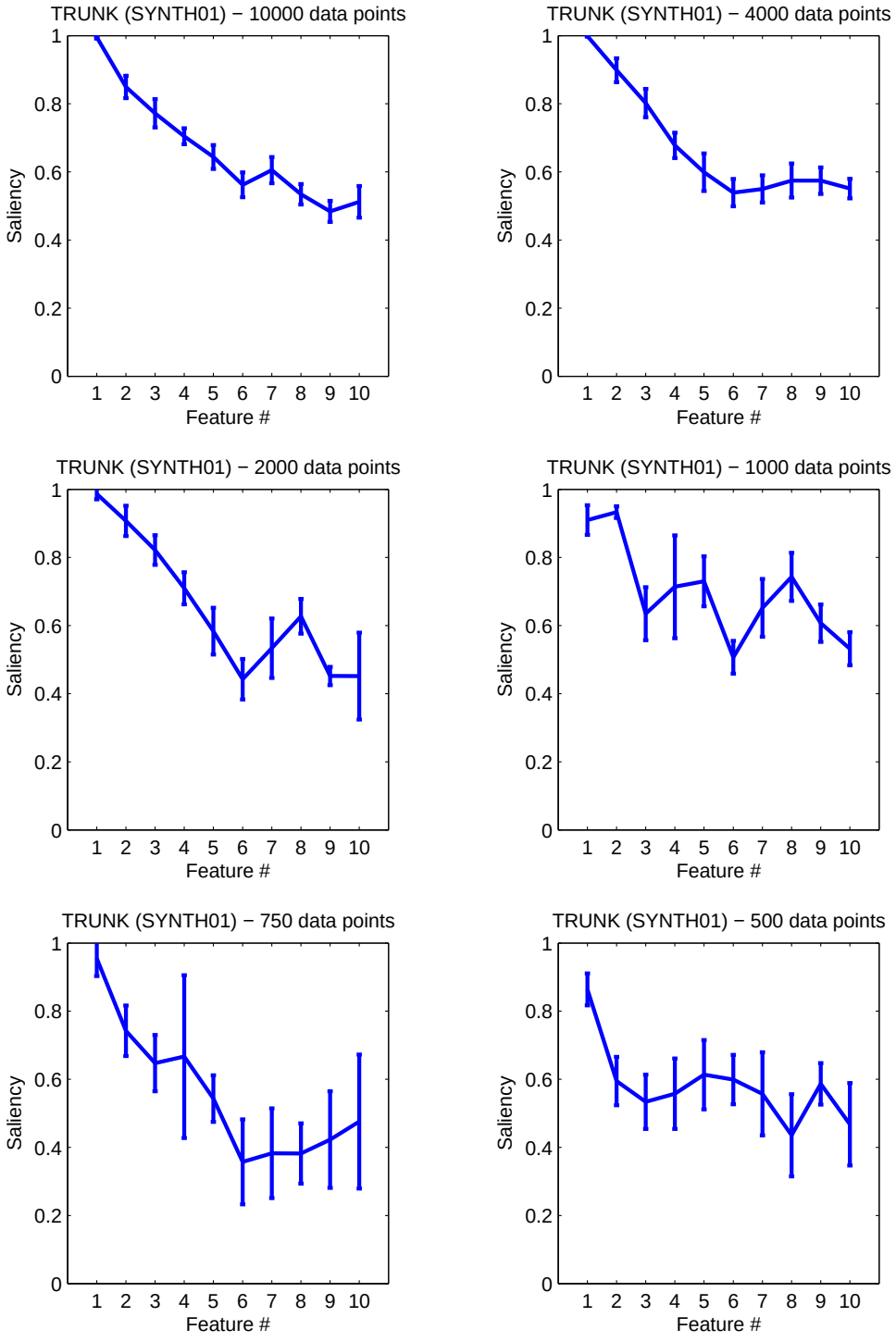

FIGURE 4.9: Experiments with different *SYNTH1* sample sizes (indicated in the plot titles) Mean saliencies ρ_d for the 10 features. The bars span from the mean minus to the mean plus one standard deviation of the saliencies over 20 runs of the algorithm.

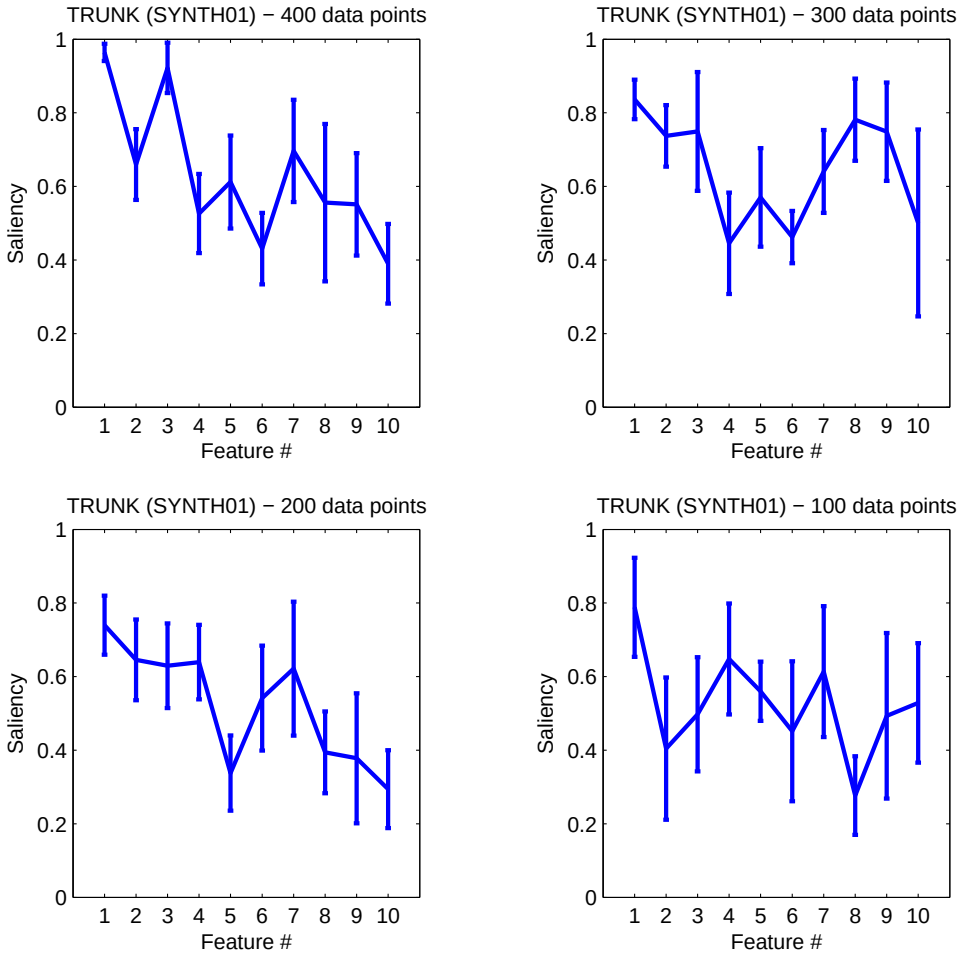

FIGURE 4.10: Experiments with different *SYNTH1* sample sizes (indicated in the plot titles) Mean saliencies ρ_d for the 10 features. The bars span from the mean minus to the mean plus one standard deviation of the saliencies over 20 runs of the algorithm (continuation of fig. 4.9).

rest of features for large sample sizes, a fatal deterioration of the saliency estimation is evident for sample sizes as big as 1,000. This partially supports *H1* but also provides a clear indication of the limitations of the technique for difficult, highly overlapping datasets.

In SYNTH4, the clusters are well separated but the more complex distributions and non-diagonal covariance matrices. The FRD ranking results for this dataset are shown in A.3 and A.4. The saliencies estimated for the two first features (and especially for the first one) are higher than those estimated for the rest of the (uninformative) features, indicating that the algorithm is perfectly capturing the more complex structure of the distributions. A deterioration of the saliency estimation is nevertheless evident for the smallest of the sample sizes investigated.

The SYNTH5 dataset consists this time of four Gaussians aligned along the first main

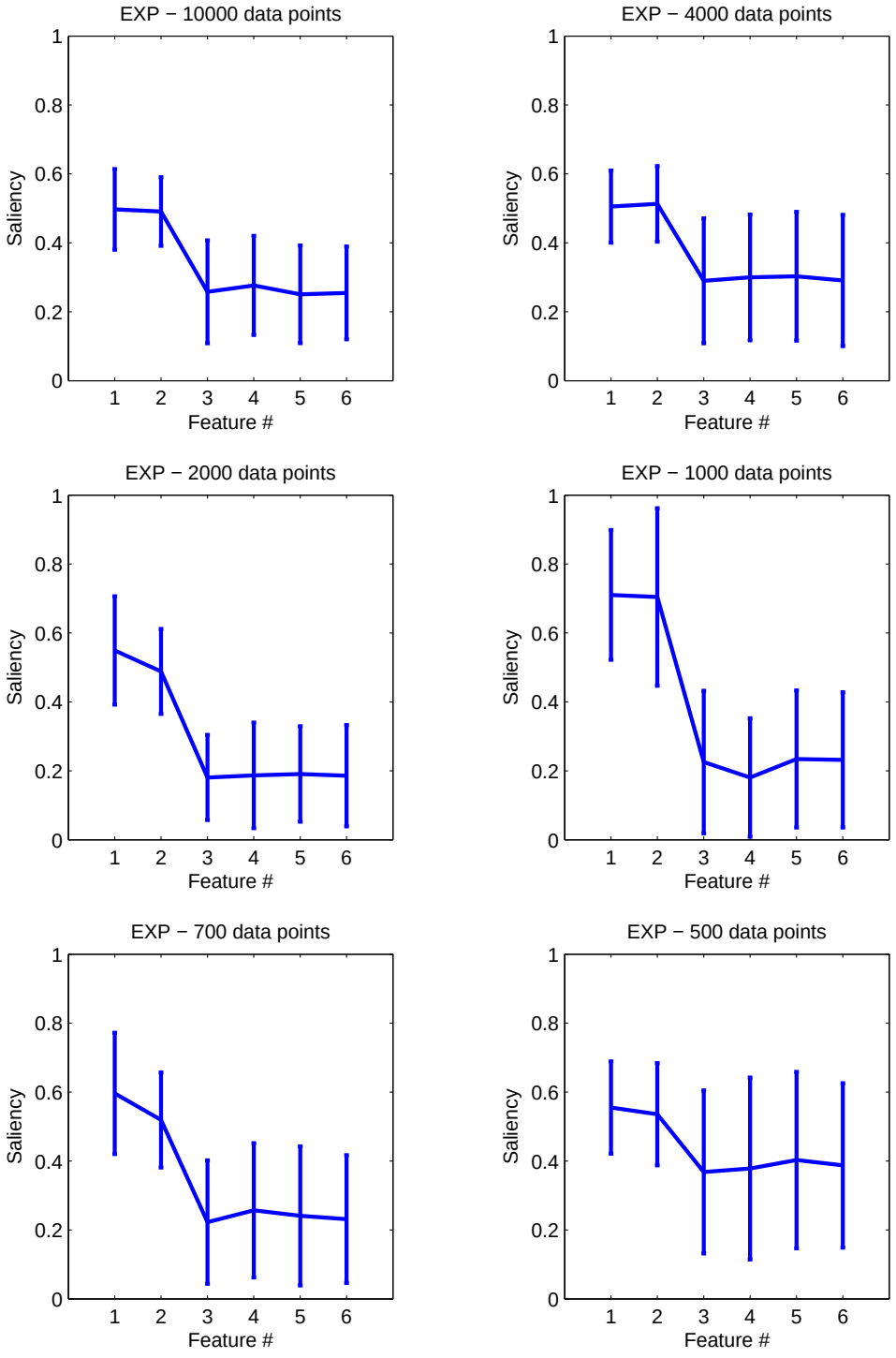

FIGURE 4.11: Experimental results for different *SYNTH2* sample sizes (indicated in the plot titles). Representation as in previous figures.

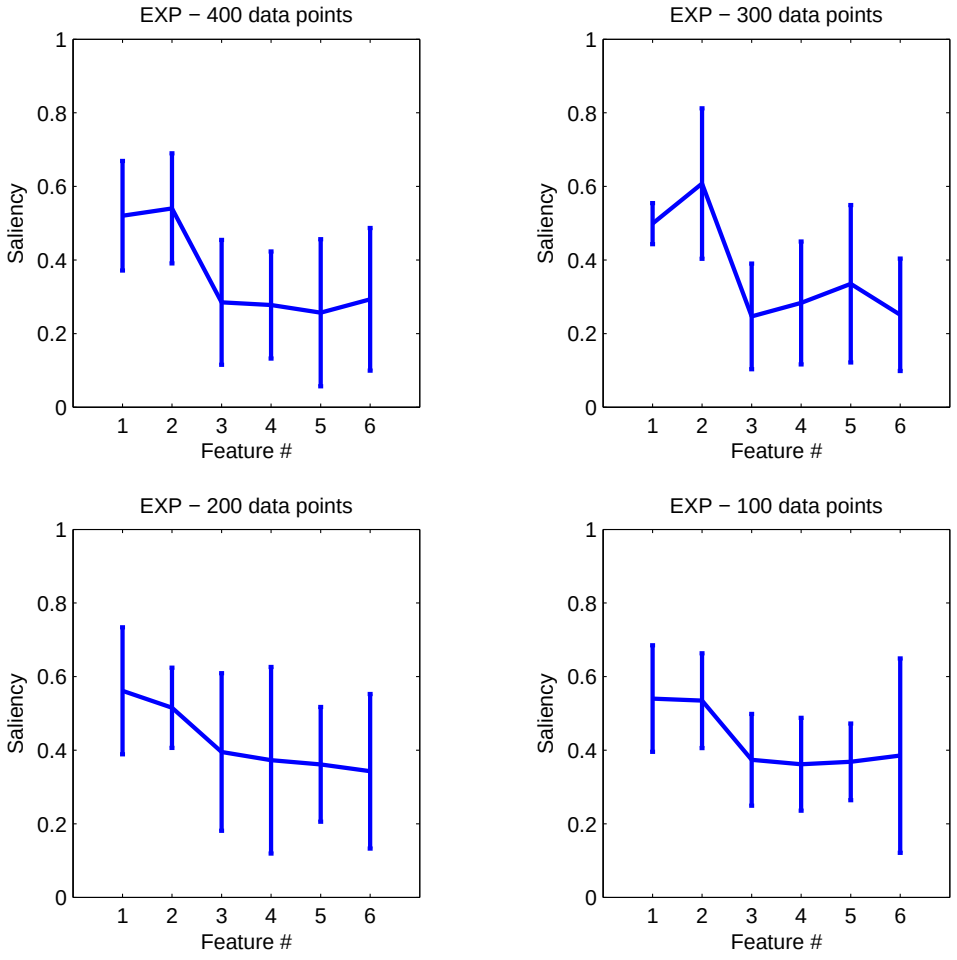

FIGURE 4.12: Experimental results for different *SYNTH2* sample sizes (continuation of fig. 4.11). Representation as in previous figures.

feature. This means that the second feature has no contribution to the overall cluster structure. The FRD ranking results for SYNTH5 are shown in Figures A.5 and A.6 and perfectly reflect the nature of these data, as the second feature consistently ranks as low as the uninformative noise features. A deterioration of the saliency estimation is again in evidence for the smallest of the sample sizes investigated, especially in the form of bigger bars of the standard deviation of the estimated saliencies.

SYNTH6, SYNTH7, SYNTH8, and SYNTH9 are variations on the same theme, and are meant to explore whether the increase in the number of clusters has any effect on FRD-GTM in terms of the sample size. SYNTH6 and SYNTH8 consist of 6 neatly defined Gaussians, whereas SYNTH7 and SYNTH9 consist of 8. SYNTH6 and SYNTH9 are arranged in a rhomboid layout, whereas SYNTH7 and SYNTH8 are arranged in a rectangular layout. The results are displayed in A.7 and A.8 (for SYNTH6), A.9 and A.10 (for SYNTH7), A.11 and A.12 (for SYNTH9), and A.13 and A.14 (for SYNTH9).

They indicate, first, that the number of Gaussians has no clear effect on saliency estimation, as the results for SYNTH6 and SYNTH9 are very similar to each other, and so are the results of SYNTH7 and SYNTH8. Secondly, they capture the modifications in the cluster structure introduced by the different layouts: for SYNTH6 and SYNTH9 the first feature is estimated to be less relevant than the second, whereas for SYNTH6 and SYNTH8 the first feature is estimated to be more relevant than the second. For all these datasets, and consistently with previous results involving well-defined clusters, the estimation of the saliency deteriorates quite gracefuly, and such deterioration is only in evidence for the smallest datasets. Overall, these results again provide partial support for hypothesis *H1*.

4.3.2 The effect of Noise

In the experiments reported in Figure 4.13, four levels of Gaussian noise of increasing level were added to a sample of 2,000 points of *SYNTH1*. The FRD-GTM is shown to behave robustly even in the presence of a substantial amount of noise, although its performance deteriorates significantly for noise of standard deviation = 1, as reflected in the breach of the expected monotonic decrease of the mean feature saliencies. It is also true that, comparing these results with those in Figures 4.9 (in which no noise was added to *SYNTH1*), the most relevant feature is not so close to a saliency of 1. *H2.1* is, therefore, partially supported by these results.

The FRD ranking results using the 10 original features of *SYNTH1* plus 5 amd 10 Gaussian noise features, are shown in Figure 4.14. For all levels of noise, the relevance (in the form of estimated saliency) of the original features ($1 \rightarrow 10$) is reasonably well estimated: the saliency for the first feature is close to 1 with almost full certainty (very small vertical bars) and, overall, the expected monotonic decrease of the mean feature saliencies is preserved, although breaches of such monotonicity can also be observed. The saliencies estimated for the 5 and 10 added Gaussian noise features are regularly estimated to be small. Interestingly, the increase in the level of noise does not seem to affect the performance of the FRD method in any significant way: the differences between the saliencies of the 10 original variables and the added (5 or 10) noisy ones stay roughly the same and the decreasing relevance for the 10 original variables does not vary substantially. According to these results, *H2.2* is not supported at this stage.

The FRD-GTM is shown to behave with reasonable robustness when noise is added to the first two features of *SYNTH2*, as shown in Figure 4.15. As in the case of *SYNTH1*, its performance deteriorates significantly for high levels of noise. Comparing these results with those in Figures 4.11 and 4.12 (in which no noise was added to the first two features),

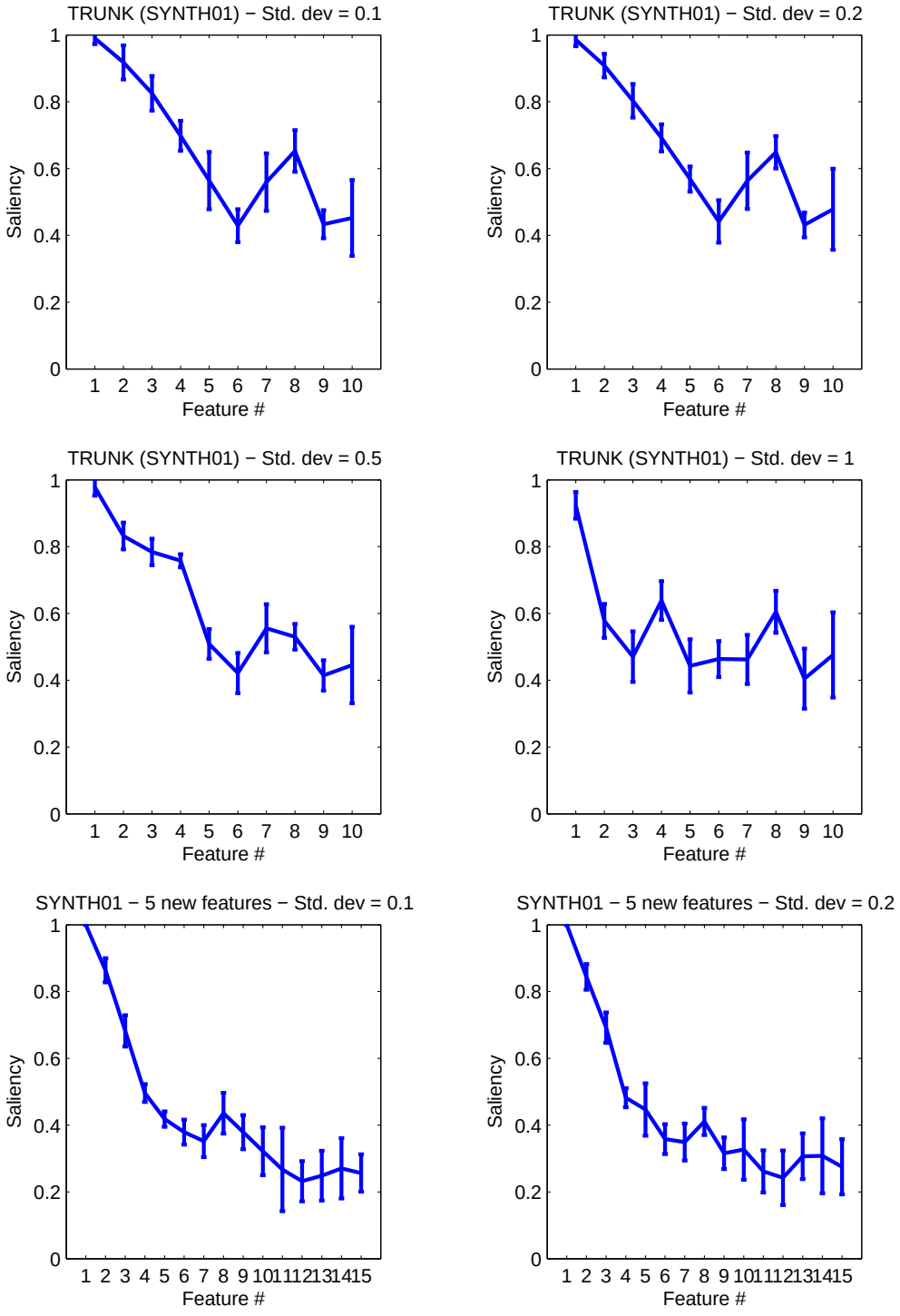

FIGURE 4.13: Experiments with a sample of 2,000 points from *SYNTH1*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

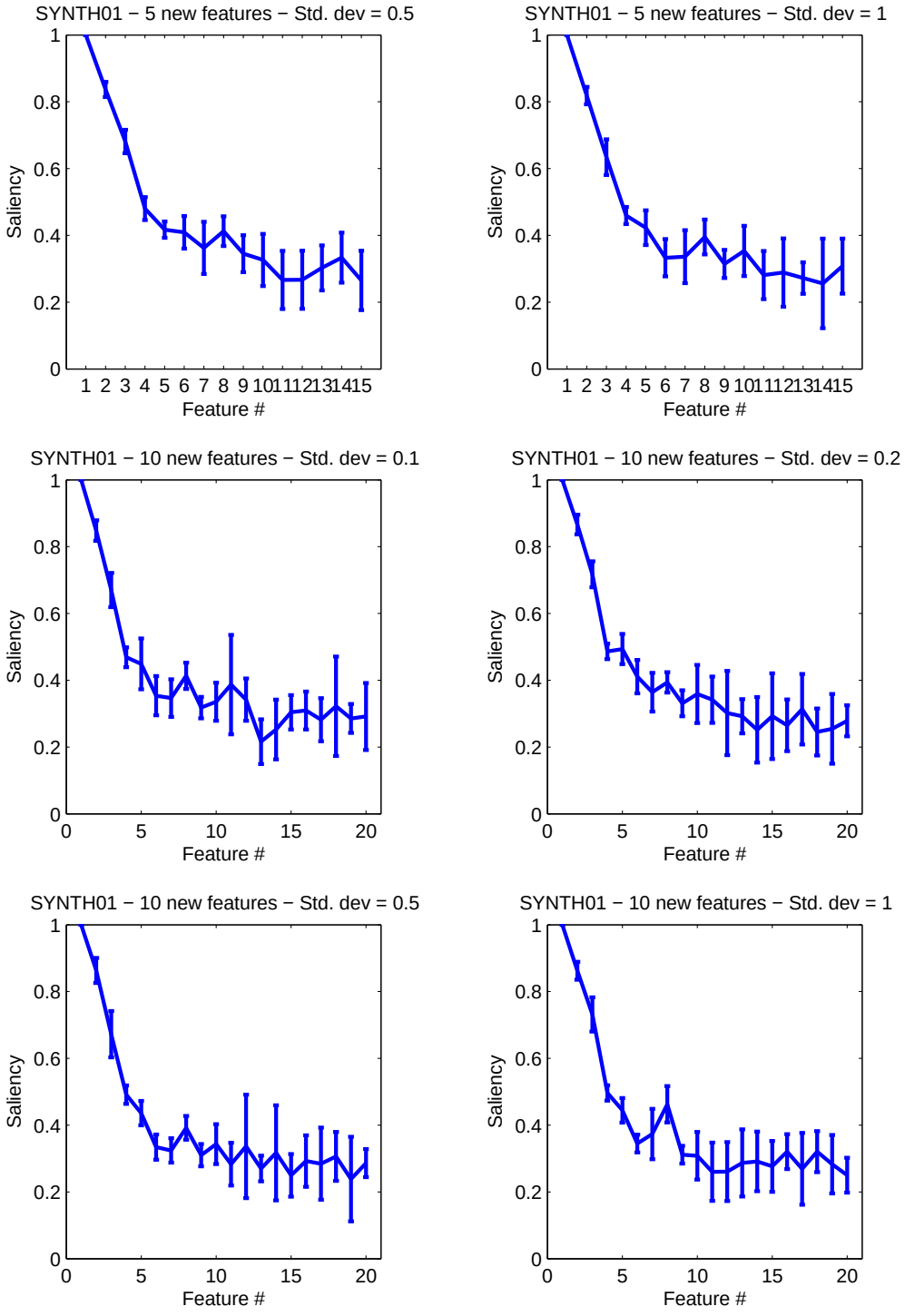

FIGURE 4.14: Experiments with a sample of 2,000 points from *SYNTH1*, to which different levels of Gaussian noise (continuation of fig. 4.13) are added. Representation as in previous figures.

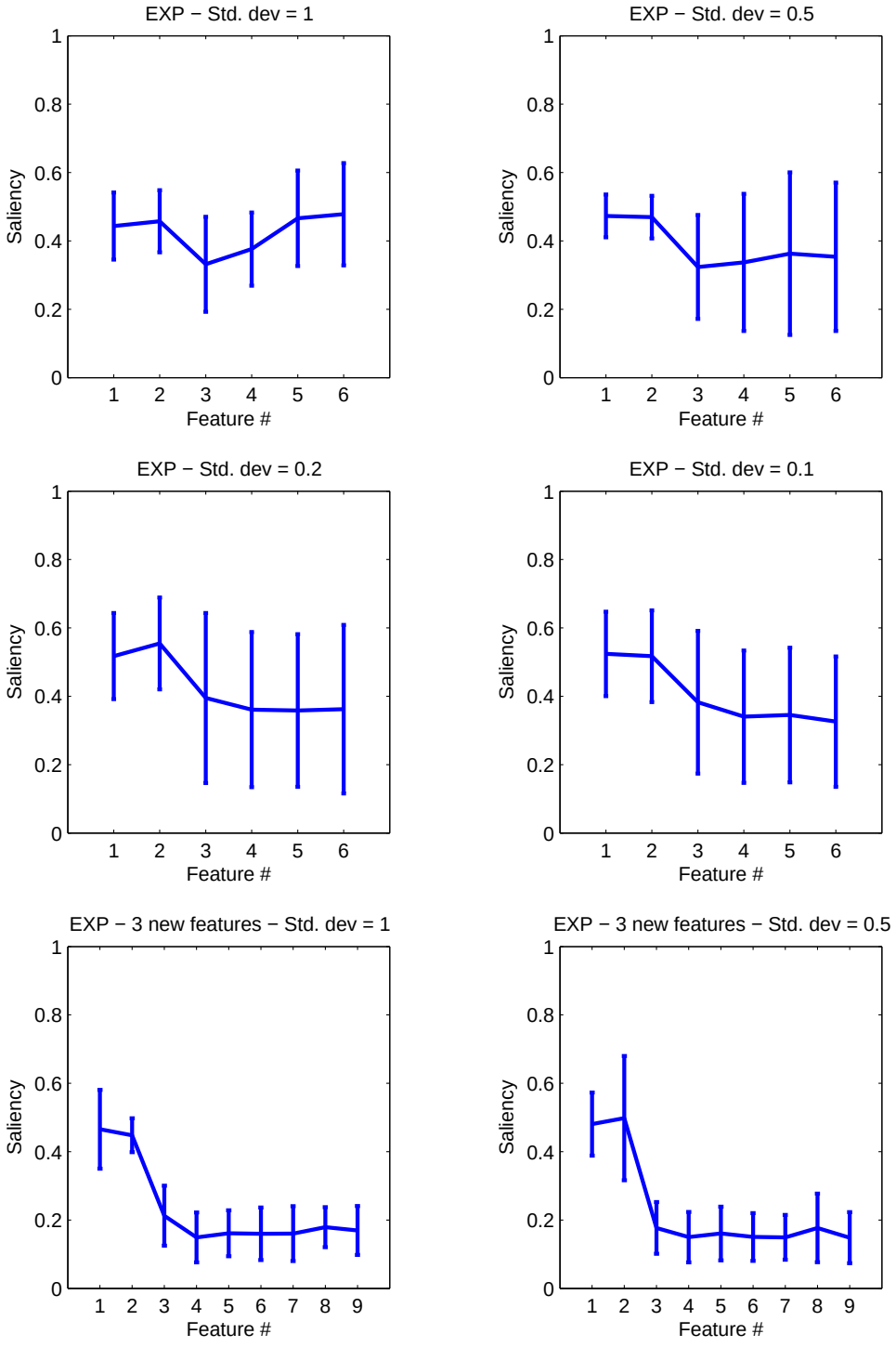

FIGURE 4.15: Experimental results for a sample size of 1000 points from *SYNTH2*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

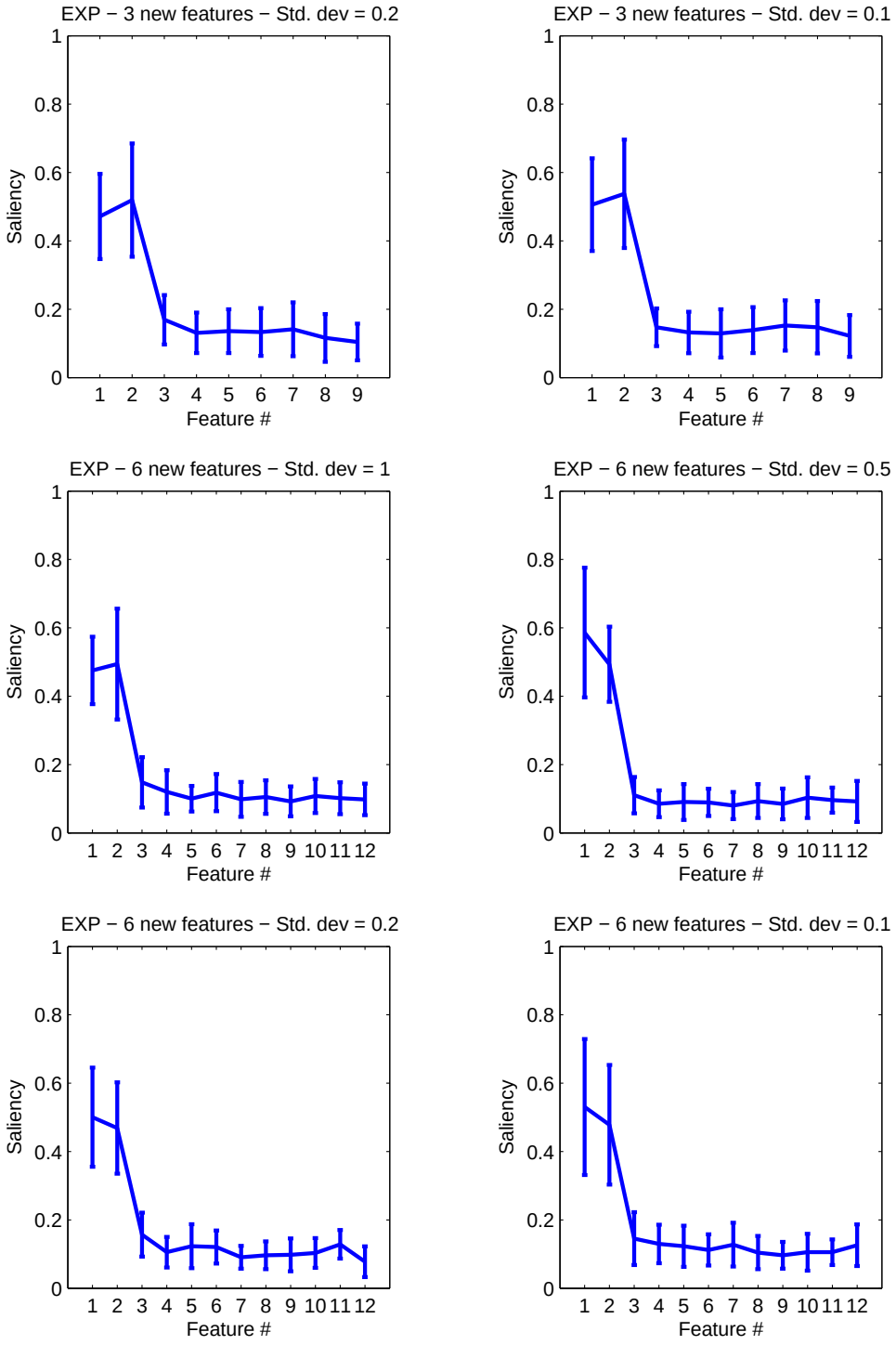

FIGURE 4.16: Experimental results for a sample size of 1000 points from *SYNTH2*, to which different levels of Gaussian noise (continuation of fig. 4.15) are added. Representation as in previous figures.

the overall deterioration becomes evident. *H2.1* is again partially supported by these results.

This support for hypothesis *H2.1* is, even if partial, certainly not unexpected. As robust as it may be, the FRD-GTM model is still prone to data overfitting. That is, at some point, the model will start learning the noise as much as learning the underlying signal distributions. The resulting FRD-GTM model will be over-complex and, if the noise is uninformative (i.e., in this case, if the noise affects all data features equally), the method of relevance determination will eventually start struggling to provide correct saliency estimations. One way around this problem is to endow the model with regularization capabilities to effectively control complexity [9–11]. FRD-GTM is thus likely to benefit from the definition of extensions of the model encompassing adaptive regularization.

The FRD ranking results for the experiments using the 2 original features of *SYNTH2* plus either 7 or 10 Gaussian noise features are shown, in turn, in Figure 4.16. This is clearly a far easier problem for the FRD method. Regardless the level of noise and the number of added noisy features, FRD-GTM consistently estimates the first 2 features to be the most relevant. Furthermore, the differences between the saliencies estimated for the first 2 features and the added (7 or 10) noisy ones stay roughly the same. In contrast with the results obtained in the experiments with *SYNTH1*, the estimated saliencies for all noisy features are low and quite similar. Our research hypothesis *H2.2* is not supported by these results.

To avoid cluttering the text with an excessive amount of figures, those corresponding to the following experiments are relegated to Appendix B. The FRD-GTM is shown to behave with reasonable robustness when noise is added to the first two features of *SYNTH3*-*SYNTH9*. As in the case of *SYNTH2*, its performance deteriorates significantly for high levels of noise. Comparing these results with those in which no noise was added to the first two features, the overall deterioration becomes evident. *H2* is again partially supported by these results.

The FRD ranking results for the experiments using the 2 original features of *SYNTH3*, *SYNTH4*, *SYNTH6*, *SYNTH7*, *SYNTH8*, *SYNTH9* plus either 7 or 10 Gaussian noise, regardless the level of noise consistently estimates the first 2 features to be the most relevant. Furthermore, the differences between the saliencies estimated for the first 2 features and the added (7 or 10) noisy ones stay roughly the same. The FRD ranking results for *SYNTH5* where the second feature has no contribution to the overall cluster structure shown in Figures B.5 and B.6 and perfectly reflect the nature of these data, as the second feature consistently ranks as low as the uninformative noise features.

4.3.3 The effect of Noise on Real Data: *Ionosphere*

The *Ionosphere* dataset described in 4.2.1.10 was also analyzed. Recall that the data consist of 34 features, structured in 17 pairs of values. Each pair is formed by the real and complex parts of the values of an autocorrelation function for a pulse number of the system signal. FRD-GTM was assessed using these data in [6], showing that all real parts had higher saliencies than their complex counterparts, meaning that the real parts describing the original signal have a richer cluster structure. In the experiments reported in Figure 4.17, four increasing levels of Gaussian noise were added to *Ionosphere*. The deterioration of the results as noise increases are evident, although the relative ordering of real vs. complex components of the feature pairs are reasonably well preserved for noise levels up to 0.5.

The FRD ranking results using the 34 original features of *Ionosphere* to which 8 or 16 Gaussian noise features are added, are shown in Figure 4.18. For all levels of noise, the relevance (in the form of estimated saliency) of the original features is reasonably well estimated. The rest of the results for this experiment can be found in the Appendix B Figure B.15. The saliencies estimated for the 8 and 16 added Gaussian noise features are consistently estimated to be small. Interestingly, the increase in the level of added noise does not affect the performance of the FRD method in any significant way: the differences between the saliencies of the 34 original variables and the added (8 or 16) noisy ones stay at roughly the same levels. According to these results, *H2* is not supported.

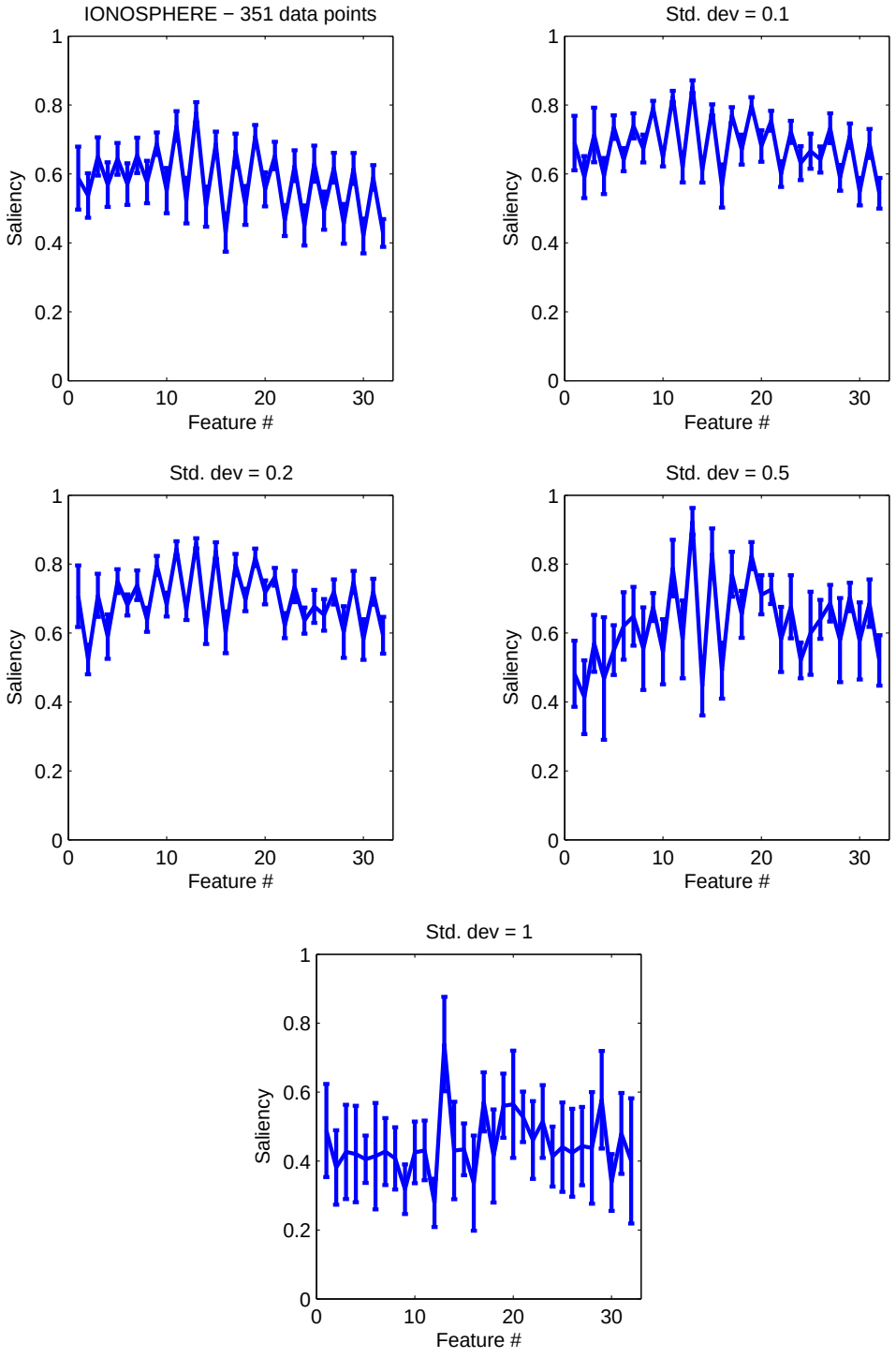

FIGURE 4.17: Experimental results for a sample size of 351 points from *IONOSPHERE*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

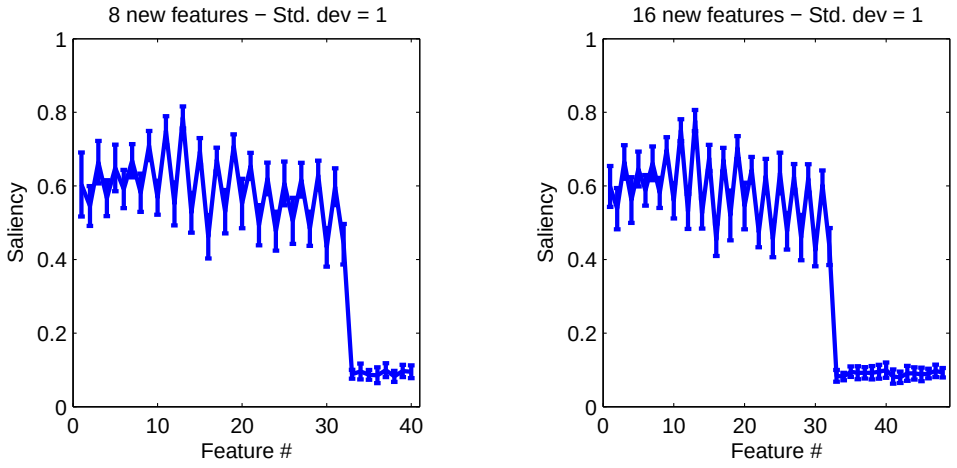

FIGURE 4.18: Experimental results for a sample size of 351 points from *IONOSPHERE*, to which different levels of Gaussian noise (continuation of fig. 4.17) are added. Representation as in previous figures.

Chapter 5

Potential Alternatives to Minimize the Impact of Noise in FRD-GTM

5.1 Introduction

In the previous chapter, we have evaluated the robustness of FRD-GTM in the presence of different levels of uninformative noise. The FRD-GTM model, in its standard version, will fit the noise indistinctly. It is therefore prone to suffer the problem of overfitting. Overfitting, as reported in the experiments, affects the feature relevance ranking, at least to some extent. In this chapter, we outline some of the potential approaches to deal with this negative effect of noise in the model's performance. They include: Firstly, some regularized variants of GTM that make use of a partially Bayesian formulation of the problem and of the evidence approach. Secondly, a full Bayesian approach to GTM training with a variational algorithmic approximation. Finally, a variation of the standard GTM that penalizes interpoint off-manifold distances while prioritizing distances along the manifold.

5.2 Regularized GTM with Feature Relevance Determination

The optimization of Eq. 3.5 does not prevent the model fitting whatever noise is present in the dataset. As mentioned elsewhere, one of the advantages of the probabilistic definition of the GTM is the possibility of introducing adaptive regularization in the mapping. This procedure automatically regulates the level of map smoothing necessary to avoid data overfitting, resorting to either a *single regularization term* (SRT) [9], or to multiple ones (in a procedure called *Selective Map Smoothing* (SMS): [10]). The first case entails the definition of a *penalized* log-likelihood of the form: $\ell_{\text{PEN}}(\mathbf{W}, \beta) = \ell(\mathbf{W}, \beta) - \frac{1}{2}\varsigma \|\mathbf{w}\|^2$, where $\ell(\mathbf{W}, \beta)$ is the log-likelihood of the original formulation of GTM (logarithm of Eq. 3.5); ς is a regularization coefficient; and $\mathbf{w}$ is a vector shaped by concatenation of the different column vectors of the weight matrix $\mathbf{W}$.

A Bayesian approach to the estimation of the regularization coefficient ς, as well as the inverse variance β, was introduced in [9]. In this procedure, Bayes' theorem is used to estimate the distributions of ς and β, given the data points, in the form:

$$p(\varsigma, \beta|\mathbf{X}) = \frac{p(\mathbf{X}|\varsigma, \beta)\, p(\varsigma, \beta)}{p(\mathbf{X})} \tag{5.1}$$

Assuming uninformative priors, the optimization of equation 5.1 is equivalent to the maximization of the *evidence*, or marginal likelihood:

$$p(\mathbf{X}|\varsigma, \beta) = \int p(\mathbf{X}|\mathbf{w}, \beta)\, p(\mathbf{w}|\varsigma)\, d\mathbf{w}, \tag{5.2}$$

for which a normal prior $p(\mathbf{w}, \varsigma) = \left(\frac{\varsigma}{2\pi}\right)^{W/2} \exp\left(-\frac{1}{2}\varsigma \|\mathbf{w}\|^2\right)$ is choosen for the weights, where W is the number of weights in matrix $\mathbf{W}$. The log-evidence or marginal log-likelihood for ς and β is then given by:

$$\ln p(\mathbf{X}|\varsigma, \beta) = \ell(\mathbf{W}_*, \beta) - \frac{1}{2}\varsigma \|\mathbf{w}_*\|^2 - \frac{1}{2}\ln|\mathbf{H}_*| + \frac{W}{2}\ln\varsigma + C \tag{5.3}$$

where $\mathbf{W}_*$ is the value of $\mathbf{w}$ in matrix form at the maximum of the posterior distribution (Eq. 5.2) and $\mathbf{H}_*$ is the Hessian of $p(\mathbf{X}|\mathbf{w}_*, \beta)\, p(\mathbf{w}_*|\varsigma)$. All the constant terms have been grouped as C. The maximization of this equation for ς and β leads to the standard updating formulae of the evidence approximation.

Alternatively, multiple regularization terms can also be considered, one for each basis function. This method, known as SMS, was originally introduced in [10]. In SMS, the prior distribution over the weights is given by

$$p\left(\mathbf{w},\{\varsigma_s\}\right)=\prod_{s=1}^{S}\left(\frac{\varsigma_s}{2\pi}\right)^{D/2}\exp\left(-\frac{1}{2}\sum_{s=1}^{S}\varsigma_s\left\|\mathbf{w}_s\right\|^2\right) \tag{5.4}$$

where each ς_s is a regularization coefficient for each basis function, and $\mathbf{w}_s$ is the vector of weights in matrix $\mathbf{W}$ that are associated with the hyperparameter s. The marginal log-likelihood of Eq. 5.3 is reformulated as:

$$\ln p\left(\mathbf{X}|\left\{\varsigma_s\right\},\beta\right)=\ell\left(\mathbf{W}_*,\beta\right)-\frac{1}{2}\sum_{s=1}^{S}\varsigma_s\left\|\mathbf{w}_{*s}\right\|^2-\frac{1}{2}\ln\left|\mathbf{H}_*\left\{\varsigma_s\right\}\right|+\frac{D}{2}\sum_{s=1}^{S}\ln\varsigma_s \tag{5.5}$$

The extension of these two regularization methods to FRD-GTM should be reasonably straightforward, as it would only entail adding the regularization terms to the likelihood expression for FRD-GTM and a differentiation with respect to the parameters of the model in the maximization M-step of the EM algorithm.

5.3 Variational Bayesian FRD-GTM

The regularization methods have been proposed in the literature [2, 10] to avoid overfitting when modelling data using GTM, described in the previous section, are based on Bayesian evidence approaches, whose efficiency is limited by some of the simplifying assumptions they require. Alternatively, we could reformulate GTM within a fully Bayesian approach and endow the model with regularization capabilities based on variational techniques. Variational inference allows approximating the marginal log-likelihood through Jensen's inequality as follows:

$$\begin{aligned}
\ln p\left(\mathbf{X}\right) &= \ln\int p\left(\mathbf{X}|\mathbf{Z},\boldsymbol{\Theta}\right)p\left(\mathbf{Z}\right)p\left(\boldsymbol{\Theta}\right)d\mathbf{Z}d\boldsymbol{\Theta}\\
&= \ln\int q\left(\mathbf{Z},\boldsymbol{\Theta}\right)\frac{p\left(\mathbf{X}|\mathbf{Z},\boldsymbol{\Theta}\right)p\left(\mathbf{Z}\right)p\left(\boldsymbol{\Theta}\right)}{q\left(\mathbf{Z},\boldsymbol{\Theta}\right)}d\mathbf{Z}d\boldsymbol{\Theta}\\
&\geq \int q\left(\mathbf{Z},\boldsymbol{\Theta}\right)\ln\frac{p\left(\mathbf{X}|\mathbf{Z},\boldsymbol{\Theta}\right)p\left(\mathbf{Z}\right)p\left(\boldsymbol{\Theta}\right)}{q\left(\mathbf{Z},\boldsymbol{\Theta}\right)}d\mathbf{Z}d\boldsymbol{\Theta}\\
&= F\left(q\left(\mathbf{Z},\boldsymbol{\Theta}\right)\right)
\end{aligned} \tag{5.6}$$

The function $F\left(q\left(\mathbf{Z},\boldsymbol{\Theta}\right)\right)$ is a lower bound such that its convergence guarantees the convergence of the marginal likelihood. The goal in variational inference is choosing a suitable form for the density $q\left(\mathbf{Z},\boldsymbol{\Theta}\right)$ in such a way that $F\left(q\right)$ can be readily evaluated

and yet which is sufficiently flexible that the bound is reasonably tight. A reasonable approximation for $q(\mathbf{Z}, \mathbf{\Theta})$ is based on the assumption that the hidden membership variables $\mathbf{Z}$ and the model parameters $\mathbf{\Theta}$ are independently distributed, i.e. $q(\mathbf{Z}, \mathbf{\Theta}) = q(\mathbf{Z})\, q(\mathbf{\Theta})$. Thereby, a Variational EM algorithm can be derived, consisting of the following basic steps:

VBE-Step:

$$q(\mathbf{Z})^{(\text{new})} \leftarrow \underset{q(\mathbf{Z})}{\operatorname{argmax}}\, F\left(q(\mathbf{Z})^{(\text{old})}, q(\mathbf{\Theta})\right) \tag{5.7}$$

VBM-Step:

$$q(\mathbf{\Theta})^{(\text{new})} \leftarrow \underset{q(\mathbf{\Theta})}{\operatorname{argmax}}\, F\left(q(\mathbf{Z})^{(\text{new})}, q(\mathbf{\Theta})\right) \tag{5.8}$$

5.3.1 Variational Bayesian EM for FRD-GTM

The steps outlined in the previous subsection are substantiated as follows

5.3.1.1 The VBE Step

$$q(\mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \tilde{\gamma}_{kn}^{z_{kn}} \tag{5.9}$$

where

$$\tilde{\gamma}_{kn} = \frac{\exp\left\{\sum_{d=1}^{D} \langle \eta_d \rangle \langle \ln p_{knd} \rangle_{\mathbf{Y},\beta} + (1 - \langle \eta_d \rangle) \langle \ln p_{0nd} \rangle_{\mathbf{y_0},\boldsymbol{\beta_0}}\right\}}{\sum_{k'=1}^{K} \exp\left\{\sum_{d=1}^{D} \langle \eta_d \rangle \langle \ln p_{knd} \rangle_{\mathbf{Y},\beta} + (1 - \langle \eta_d \rangle) \langle \ln p_{0nd} \rangle_{\mathbf{y_0},\boldsymbol{\beta_0}}\right\}} \tag{5.10}$$

5.3.1.2 The VBM Step

The variational distribution $q(\mathbf{\Theta})$ can be approximated to the product of the variational distribution of each one of the parameters if they are assumed to be independent and identically distributed. If so, $q(\mathbf{\Theta})$ is expressed as:

$$q(\mathbf{\Theta}) = q(\mathbf{Y})\, q(\beta)\, q(\boldsymbol{\eta})\, q(\mathbf{y_0})\, q(\boldsymbol{\beta_0}) \tag{5.11}$$

where natural choices of $q(\mathbf{Y})$, $q(\beta)$, $q(\boldsymbol{\eta})$, $q(\mathbf{y_0})$ and $q(\boldsymbol{\beta_0})$ are similar distributions to the priors $p(\mathbf{Y})$, $p(\beta)$, $p(\boldsymbol{\eta})$, $p(\mathbf{y_0})$ and $p(\boldsymbol{\beta_0})$, respectively. Thus,

$$q\left(\mathbf{Y}\right) = \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{y}_{(d)} | \tilde{\mathbf{m}}^{(d)}, \tilde{\mathbf{\Sigma}}^{(d)}\right) \tag{5.12}$$

$$q\left(\beta\right) = \Gamma\left(\beta | \tilde{d}_{\beta}, \tilde{s}_{\beta}\right) \tag{5.13}$$

$$q\left(\boldsymbol{\eta}\right) = \prod_{d=1}^{D} \tilde{\rho}_{d}{}^{\eta_{d}} \tag{5.14}$$

$$q\left(\mathbf{y_0}\right) = \prod_{d=1}^{D} \mathcal{N}\left(y_{0d} | \tilde{m}_{0d}, \tilde{\tau}_{0d}\right) \tag{5.15}$$

$$q\left(\boldsymbol{\beta_0}\right) = \prod_{d=1}^{D} \Gamma\left(\beta_{0d} | \tilde{d}_{\beta_{0d}}, \tilde{s}_{\beta_{0d}}\right) \tag{5.16}$$

Then, the variational parameters are estimated to be:

$$\tilde{\mathbf{\Sigma}}_{(d)} = \left(\left\langle\beta\right\rangle \left\langle\eta_{d}\right\rangle \sum_{n=1}^{N} \mathbf{G}_{n} + \mathbf{C}^{-1}\right)^{-1} \tag{5.17}$$

$$\tilde{\mathbf{m}}^{(d)} = \left\langle\beta\right\rangle \left\langle\eta_{d}\right\rangle \tilde{\mathbf{\Sigma}} \sum_{n=1}^{N} x_{nd} \left\langle\mathbf{z}_{n}\right\rangle \tag{5.18}$$

$$\tilde{d}_{\beta} = d_{\beta} + \frac{N}{2} \sum_{d=1}^{D} \left\langle\eta_{d}\right\rangle \tag{5.19}$$

$$\tilde{s}_{\beta} = s_{\beta} + \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \left\langle z_{kn}\right\rangle \sum_{d=1}^{D} \left\langle\eta_{d}\right\rangle \left\langle \left(x_{nd} - y_{kd}\right)^{2} \right\rangle \tag{5.20}$$

$$\tilde{\rho}_{d} = \rho_{d} \exp\left\{ \sum_{n=1}^{N} \sum_{k=1}^{K} \left\langle z_{kn}\right\rangle \left[\left\langle \ln p_{knd} \right\rangle_{\mathbf{Y},\beta} - \left\langle \ln p_{0nd} \right\rangle_{\mathbf{y_0},\boldsymbol{\beta_0}} \right] \right\} \tag{5.21}$$

$$\tilde{\tau}_{0d} = N\left(1 - \left\langle\eta_{d}\right\rangle\right) \left\langle\beta_{0d}\right\rangle + \tau_{0d} \tag{5.22}$$

$$\tilde{m}_{0d} = \frac{1}{\tilde{\tau}_{0d}} \left[\left(1 - \left\langle\eta_{d}\right\rangle\right) \left\langle\beta_{0d}\right\rangle \sum_{n=1}^{N} x_{nd} + \tau_{0d} m_{0d} \right] \tag{5.23}$$

Algorithm 1 Variational algorithm

Step 1 Initialize model parameters
repeat:

Step 2 VBE Step

Step 3 VBM Step

until convergence is guaranteed

$$\tilde{d}_{\beta_{0d}} = \frac{N}{2}\left(1 - \langle \eta_d \rangle\right) + d_{\beta_{0d}} \tag{5.24}$$

$$\tilde{s}_{\beta_{0d}} = \frac{1}{2}\left(1 - \langle \eta_d \rangle\right) \sum_{n=1}^{N} \left(x_{nd} + \langle y_{0d} \rangle\right)^2 + \frac{N}{2}\left(1 - \langle \eta_d \rangle\right) \tilde{\tau}_{0d} + s_{\beta_{0d}} \tag{5.25}$$

Details of these calculations can be found in [20]. In a nutshell, the complete algorithm is summarized in Algorithm 1.

5.3.2 Lower Bound

We can now go back to the calculation of the lower bound expression from Eq. 5.6:

$$\begin{aligned} F(q) \;=\; & \langle \ln p\left(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta, \boldsymbol{\eta}, \mathbf{y_0}, \boldsymbol{\beta_0}\right) \rangle_{\mathbf{Z},\mathbf{Y},\beta,\boldsymbol{\eta},\mathbf{y_0},\boldsymbol{\beta_0}} \\ & - D_{KL}\left[q\left(\mathbf{Z}\right)||p\left(\mathbf{Z}\right)\right] - D_{KL}\left[q\left(\mathbf{Y}\right)||p\left(\mathbf{Y}\right)\right] - D_{KL}\left[q\left(\beta\right)||p\left(\beta\right)\right] \\ & - D_{KL}\left[q\left(\boldsymbol{\eta}\right)||p\left(\boldsymbol{\eta}\right)\right] - D_{KL}\left[q\left(\mathbf{y_0}\right)||p\left(\mathbf{y_0}\right)\right] - D_{KL}\left[q\left(\boldsymbol{\beta_0}\right)||p\left(\boldsymbol{\beta_0}\right)\right] \end{aligned} \tag{5.26}$$

where:

$$\begin{aligned} \langle \ln p\left(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta, \boldsymbol{\eta}, \mathbf{y_0}, \boldsymbol{\beta_0}\right) \rangle_{\mathbf{Z},\mathbf{Y},\beta,\boldsymbol{\eta},\mathbf{y_0},\boldsymbol{\beta_0}} \;=\; & \sum_{n=1}^{N}\sum_{k=1}^{K} \langle z_{kn} \rangle \sum_{d=1}^{D} \Big[\langle \eta_d \rangle \langle \ln p_{knd} \rangle_{\mathbf{Y},\beta} \\ & + \left(1 - \langle \eta_d \rangle\right) \langle \ln p_{0nd} \rangle_{\mathbf{y_0},\boldsymbol{\beta_0}} \Big] \end{aligned} \tag{5.27}$$

The operator $D_{KL}\left[q||p\right]$ is the Kullback-Leibler divergence between q and p.

5.4 Geodesic GTM with Feature Relevance Determination

The Geo-GTM model is an extension of GTM that favours the similarity of points along the learned manifold, while penalizing the similarity of points that are not contiguous in the manifold, even if close in terms of the Euclidean distance. This is achieved by modifying the standard calculation of the responsibilities in proportion to the discrepancy between the geodesic (approximated by a graph calculation) and the Euclidean distances. Such discrepancy is made operational through the definition of the exponential distribution

$$\mathcal{E}(d_g|d_e,\alpha) = \frac{1}{\alpha}\exp\left\{-\frac{d_g(\mathbf{x}_n,\mathbf{y}_m) - d_e(\mathbf{x}_n,\mathbf{y}_m)}{\alpha}\right\}, \tag{5.28}$$

where $d_e(\mathbf{x}_n,\mathbf{y}_m)$ and $d_g(\mathbf{x}_n,\mathbf{y}_m)$ are, in turn, the Euclidean and graph distances between data point $\mathbf{x}_n$ and the GTM prototype $\mathbf{y}_m$. The responsibilities of the model are redefined as:

$$z_{mn}^{geo} = p(\mathbf{u}_m|\mathbf{x}_n,\mathbf{W},\beta) = \frac{p'(\mathbf{x}_n|\mathbf{u}_m,\mathbf{W},\beta)p(\mathbf{u}_m)}{\sum_{m'} p'(\mathbf{x}_n|\mathbf{u}_{m'},\mathbf{W},\beta)p(\mathbf{u}_{m'})}, \tag{5.29}$$

where $p'(\mathbf{x}_n|\mathbf{u}_m,\mathbf{W},\beta) = \mathcal{N}(\mathbf{y}(\mathbf{u}_m,\mathbf{W}),\beta)\mathcal{E}(d_g(\mathbf{x}_n,\mathbf{y}_m)^2|d_e(\mathbf{x}_n,\mathbf{y}_m)^2,1)$. When there is no agreement between the graph approximation of the geodesic distance and the Euclidean distance, the value of the numerator of the fraction within the exponential in (5.28) increases, pushing the exponential and, as a result, the modified responsibility, towards smaller values, i.e., punishing the discrepancy between metrics. Once the responsibility is calculated in the modified E-step, the rest of the model's parameters are estimated following the standard EM procedure [21].

Notice that this means that points which are off-manifold will be penalized. This is what usually happens in the presence of noise. Some preliminary results using Geo-GTM [22] show that it recovers the underlying data generators far better than the standard GTM counterpart in the presence of increasing levels of noise. Again, the implementation of a Geodesic variation of FRD-GTM would be straightforward, as it would only entail a minor modification of the E-step in the EM algorithm.

5.5 Conclusions

Several methods to deal with one of the main problems analyzed in this thesis, namely the effect of uninformative noise in the performance of the FRD-GTM model, have been briefly outlined in this chapter. The theory of variational FRD-GTM, in particular, has been developed in some detail. The implementation of the other methods should be reasonably straightforward.

Both the full development of these methods, and its comparative assessment through detailed experimentation are beyond the scope of this thesis and should be targets for future research.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The effects of sample size and the presence of noise on a method of unsupervised feature relevance determination for the manifold learning GTM model have been investigated in some detail. The FRD-GTM has been shown to behave with reasonable robustness even at small sample sizes and in the presence of a fair amount of noise. Even though, performance deterioration has been observed at very small sample sizes and in the presence of high levels of noise. Overall, hypotheses *H1* and *H2.1* have been partially supported, while hypothesis *H2.2* has not been supported at all by the experimental evidence.

The relative weakness of the method in the presence of noise makes it convenient to consider possible strategies for model regularization and, therefore, future research will be devoted the design of methods for automatic and proactive model regularization to prevent or at least limit the negative effect of data overfitting on the FRD method for GTM. Some of such methods have already been designed for the standard GTM formulation [9, 10] and could be extended to FRD-GTM. Alternatively, regularization could be accomplished through a reformulation of the GTM within a variational Bayesian theoretical framework [12]. Again, this could be extended to accomodate FRD as exemplified by the theoretical development summarized in section 5.2.

6.2 Future Work

Future research should also extend the current experimental design to include a wider variety of artificial data sets of different characteristics, as well as to include comparisons with alternative unsupervised feature relevance determination and feature selection techniques.

It should also address the design of strategies for adaptive model regularization for FRD-GTM. Such kind of strategy would automatically regulate the level of map smoothing necessary to avoid the model fitting the noise in the data, i.e. data overfitting.

Appendix A

Figures: Sample Size effect

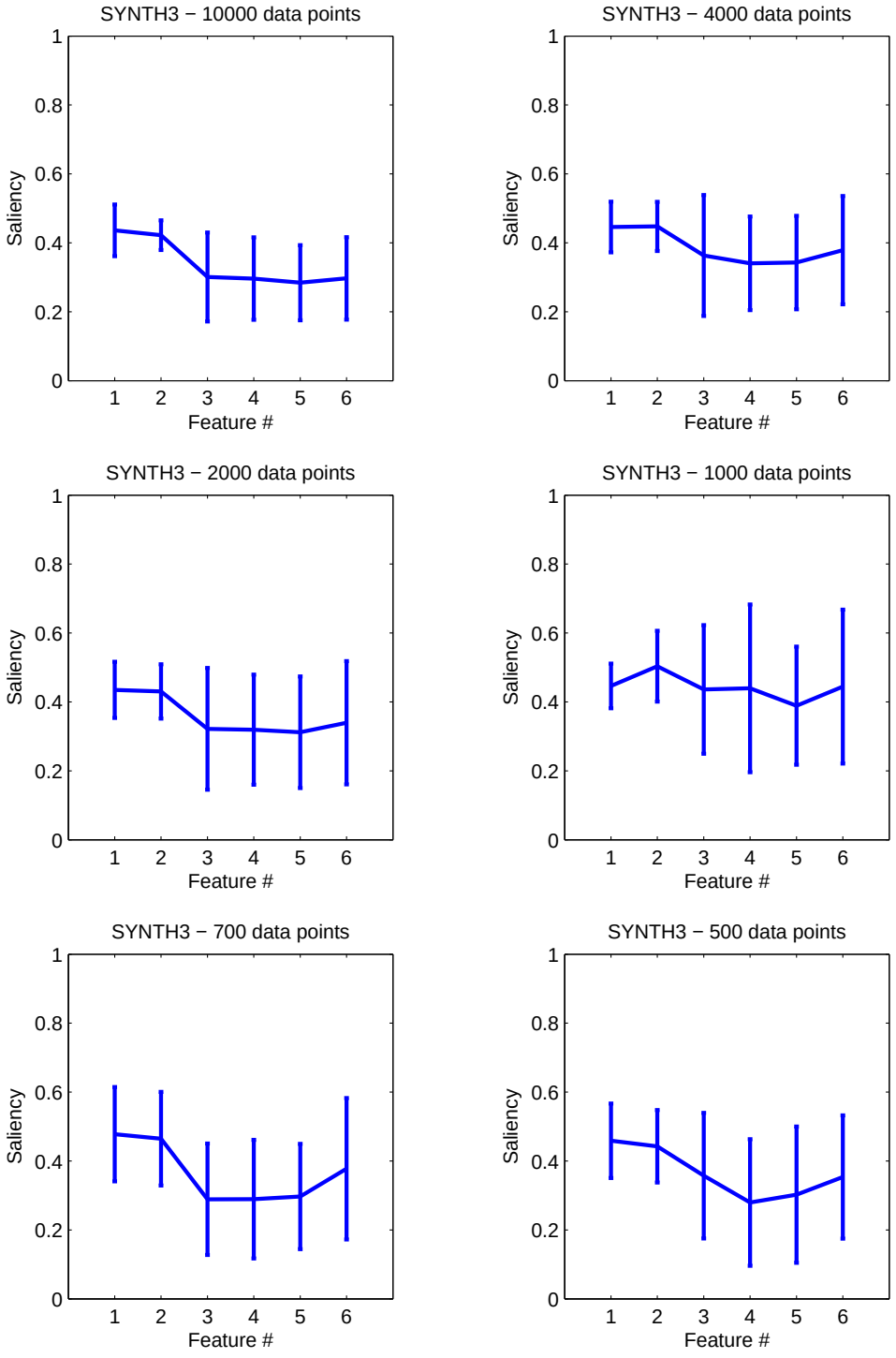

FIGURE A.1: Experimental results for different *SYNTH3* sample sizes (indicated in the plot titles). Representation as in previous figures.

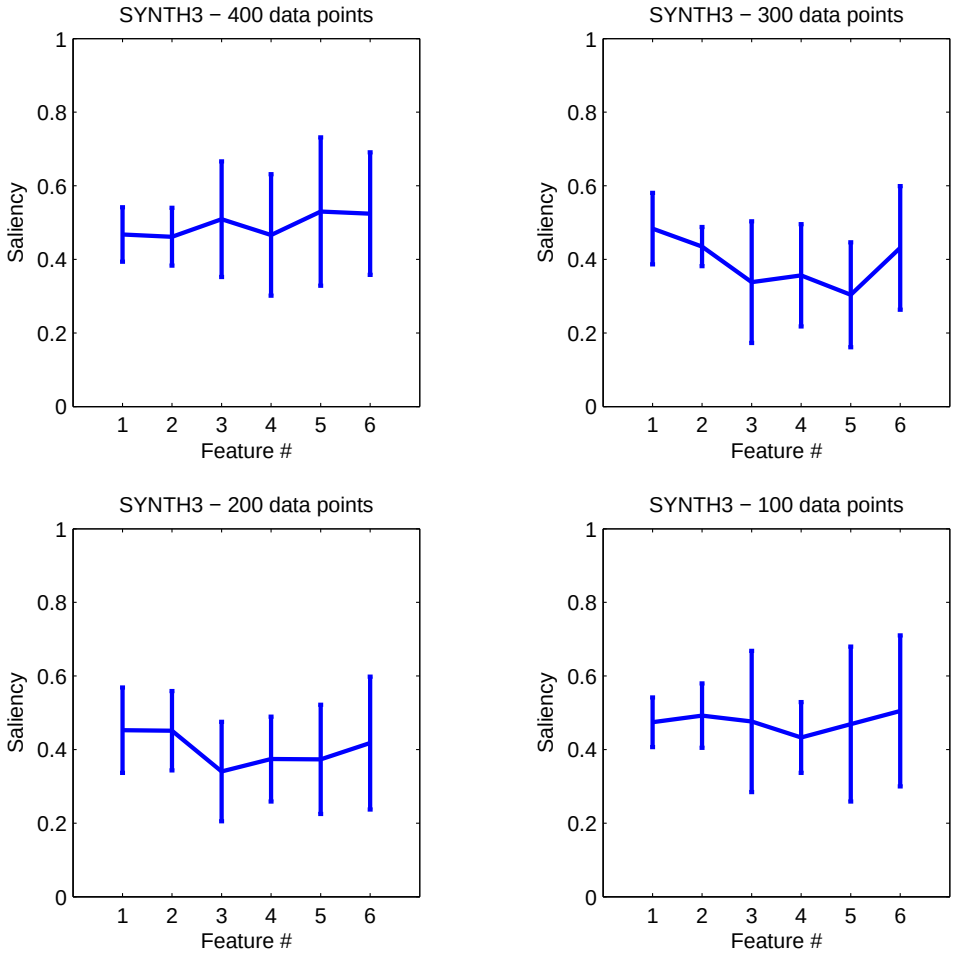

FIGURE A.2: Experimental results for different *SYNTH3* sample sizes (continuation of fig. A.1). Representation as in previous figures.

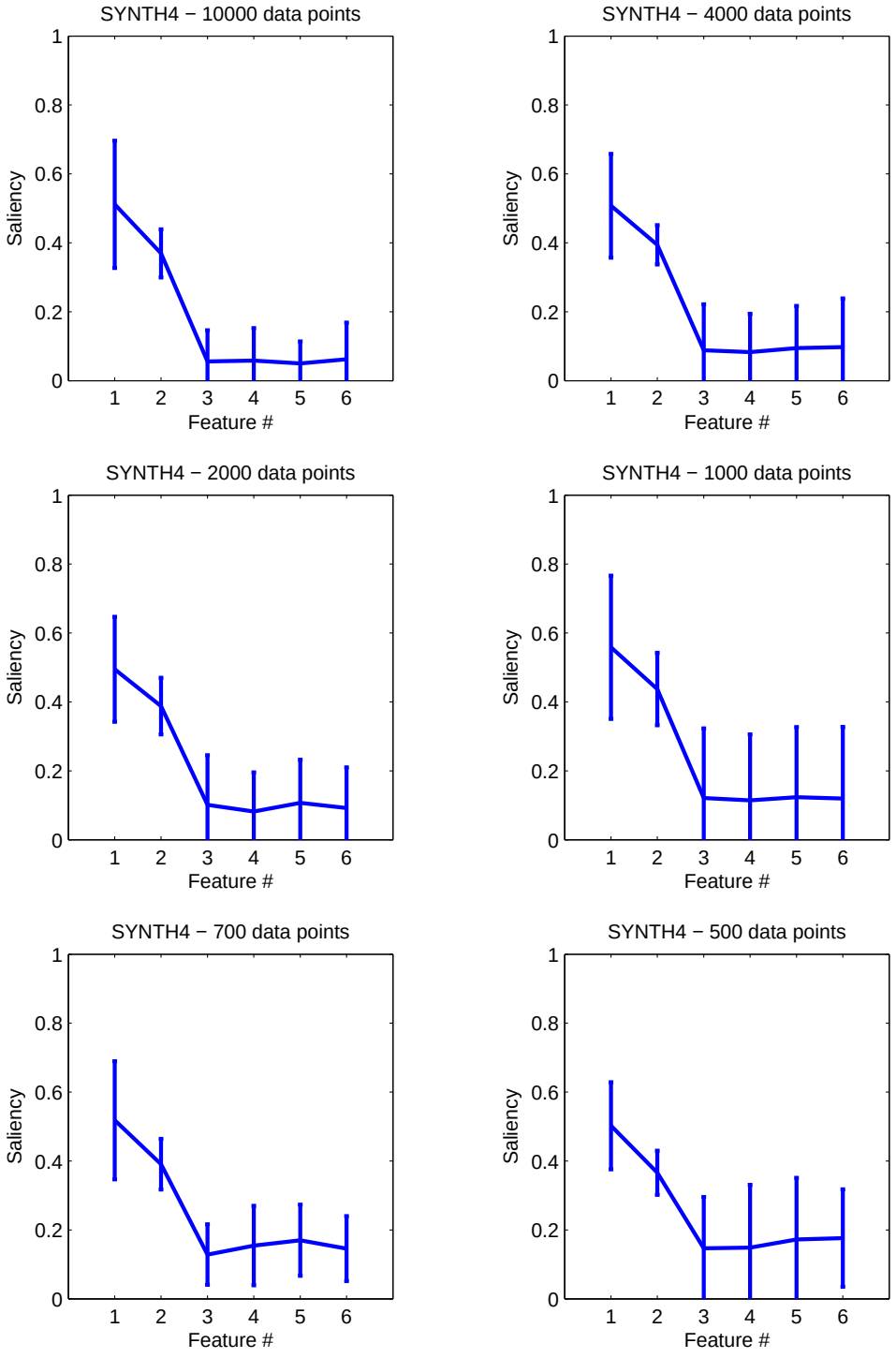

FIGURE A.3: Experimental results for different *SYNTH4* sample sizes (indicated in the plot titles). Representation as in previous figures.

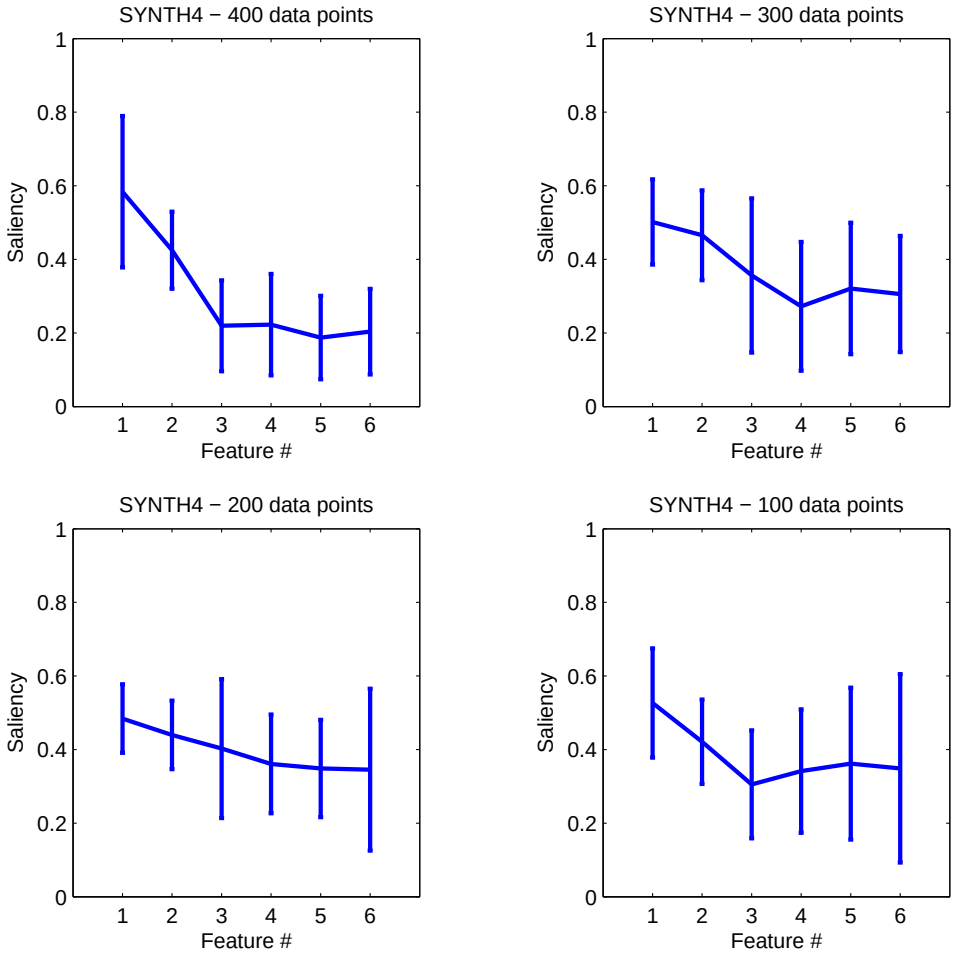

FIGURE A.4: Experimental results for different *SYNTH4* sample sizes (continuation of fig. A.3). Representation as in previous figures.

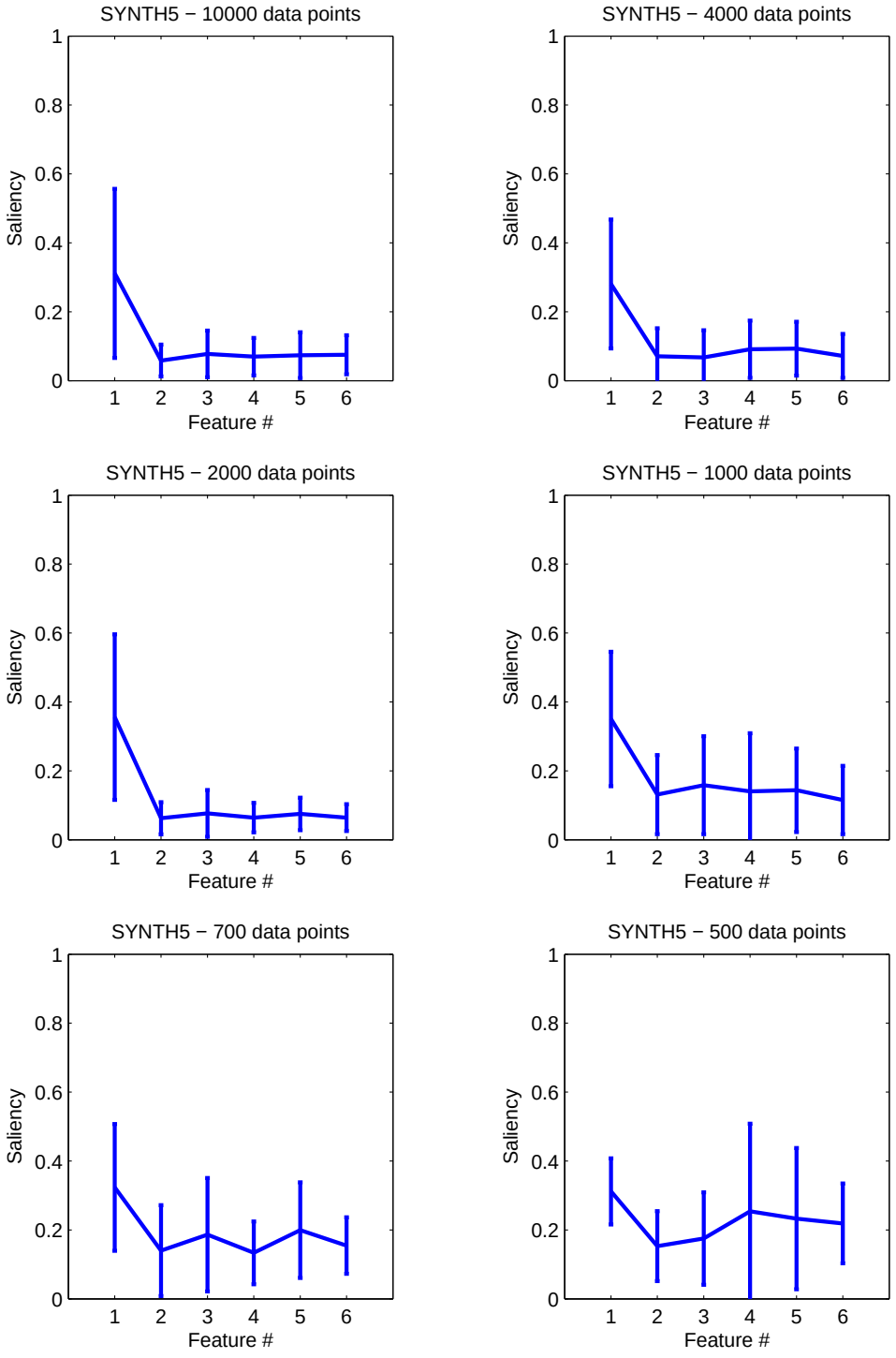

FIGURE A.5: Experimental results for different *SYNTH5* sample sizes (indicated in the plot titles). Representation as in previous figures.

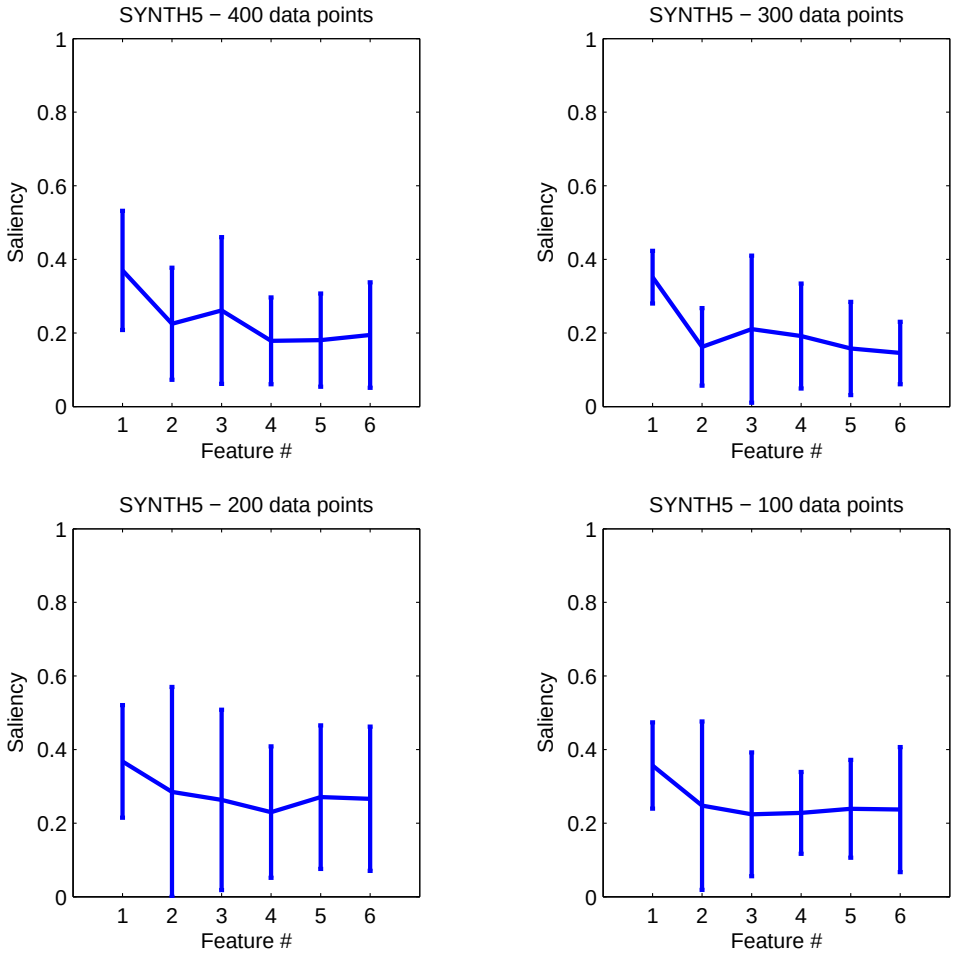

FIGURE A.6: Experimental results for different *SYNTH5* sample sizes (continuation of fig. A.5). Representation as in previous figures.

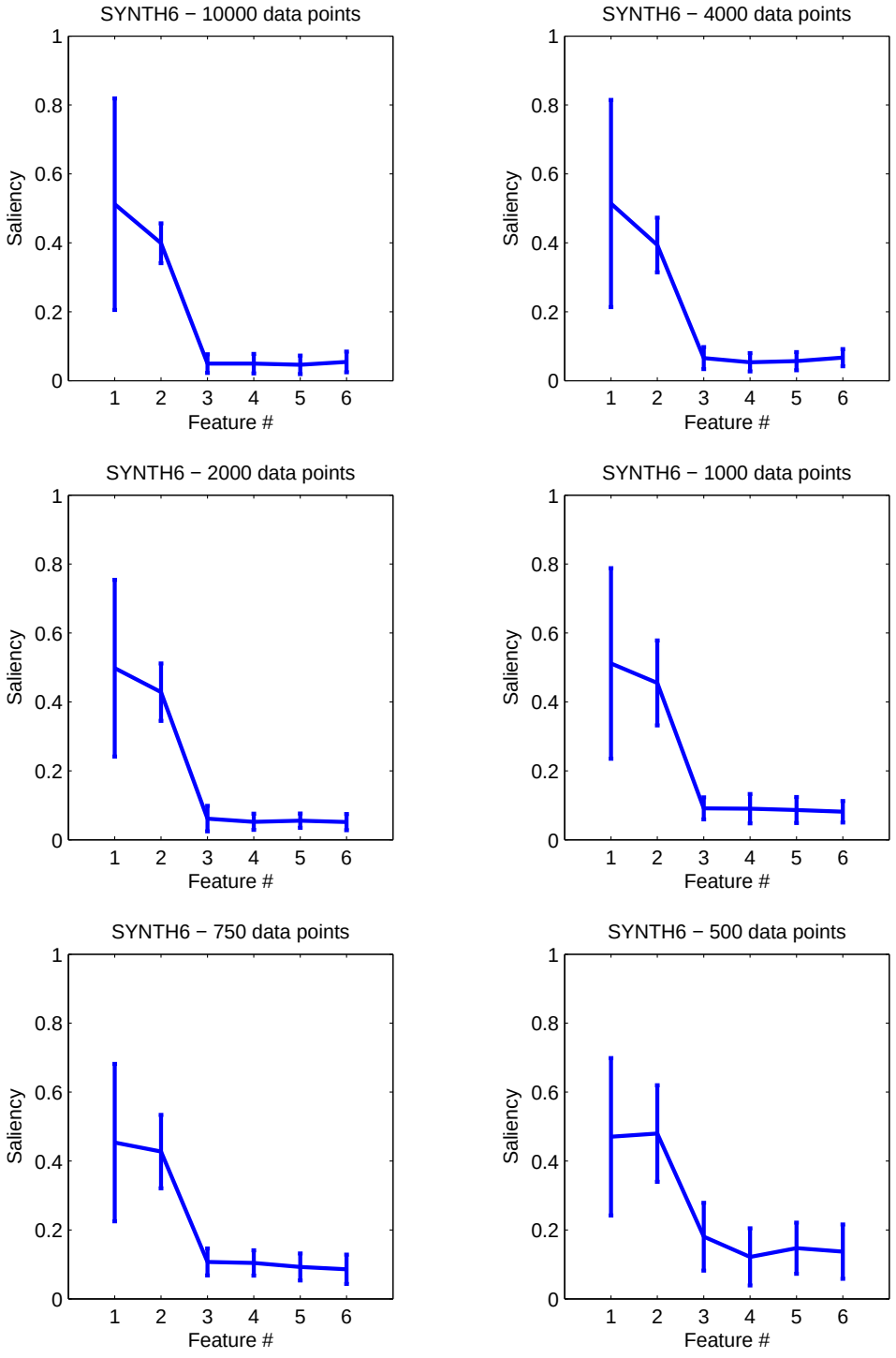

FIGURE A.7: Experimental results for different *SYNTH6* sample sizes (indicated in the plot titles). Representation as in previous figures.

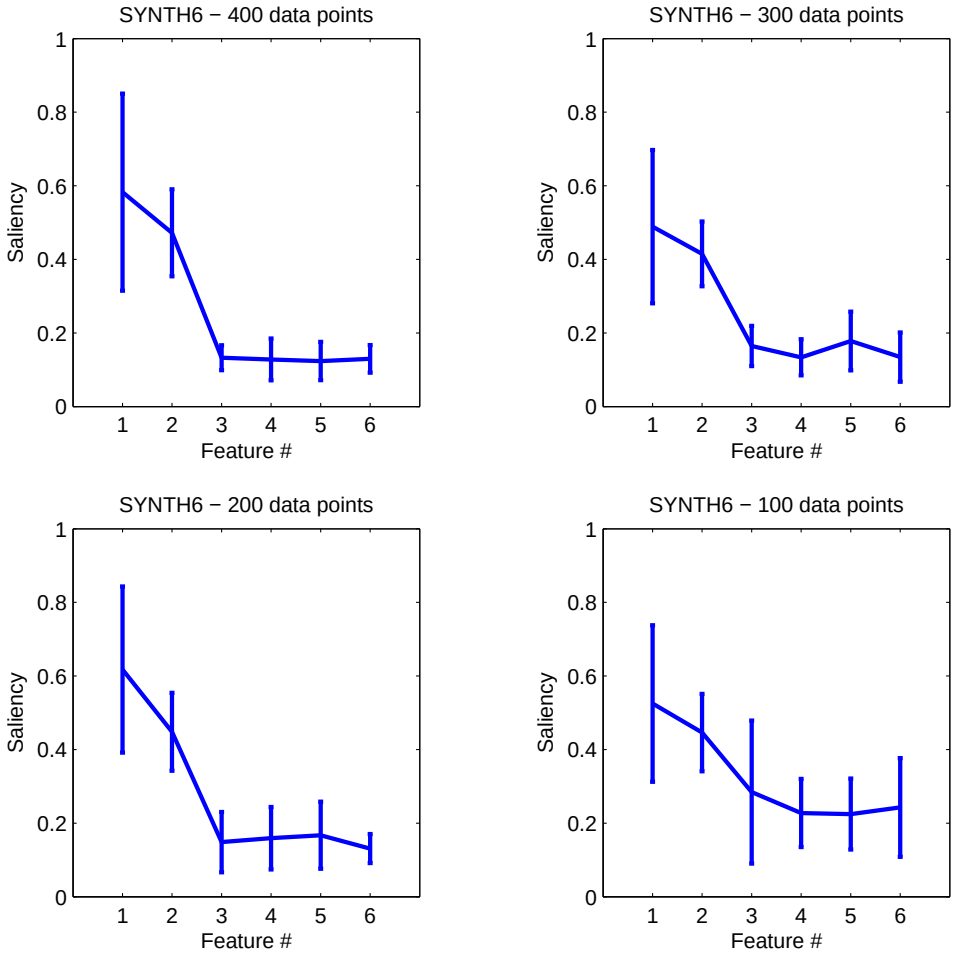

FIGURE A.8: Experimental results for different *SYNTH6* sample sizes (continuation of fig. A.7). Representation as in previous figures.

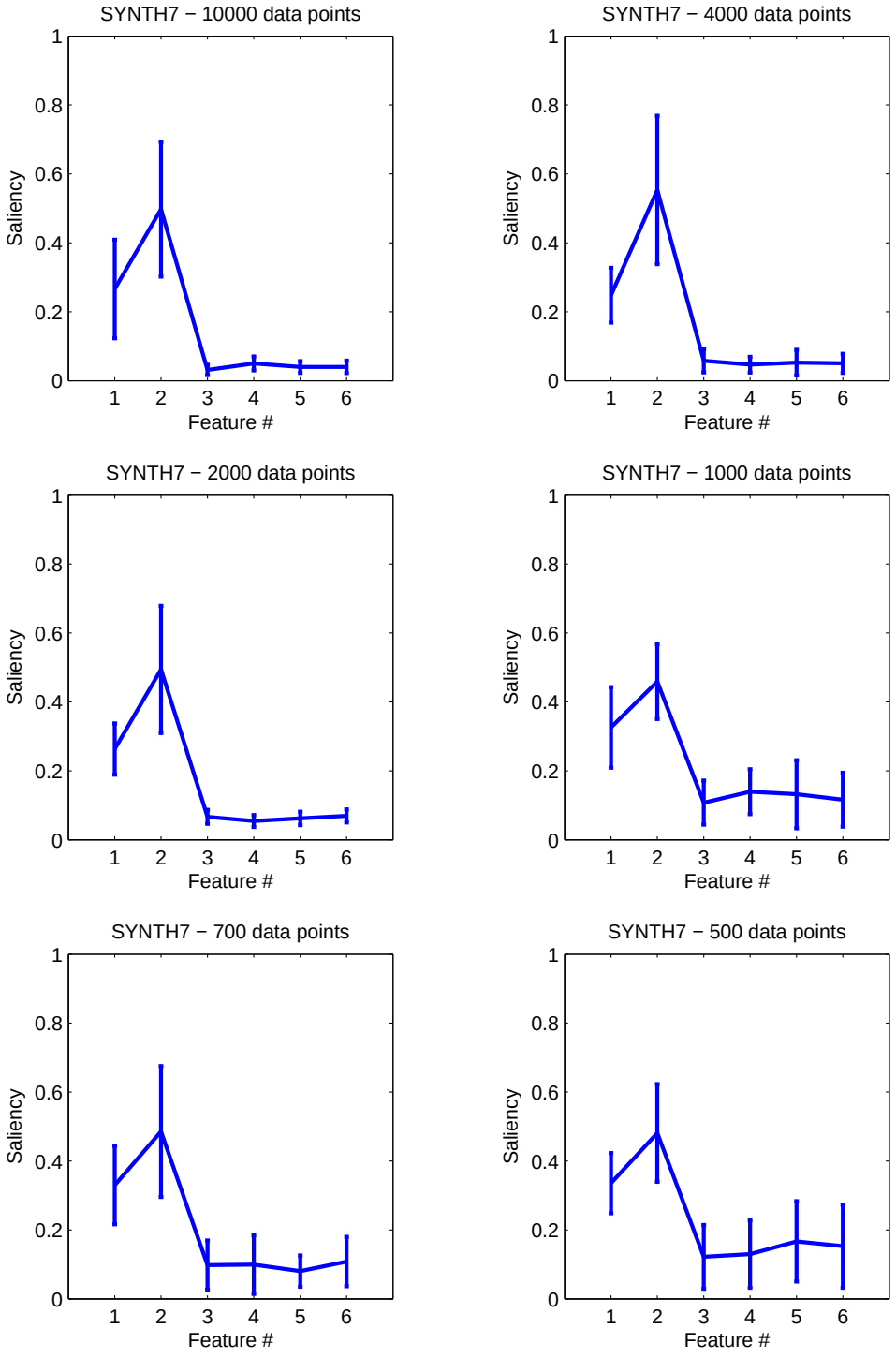

FIGURE A.9: Experimental results for different *SYNTH7* sample sizes (indicated in the plot titles). Representation as in previous figures.

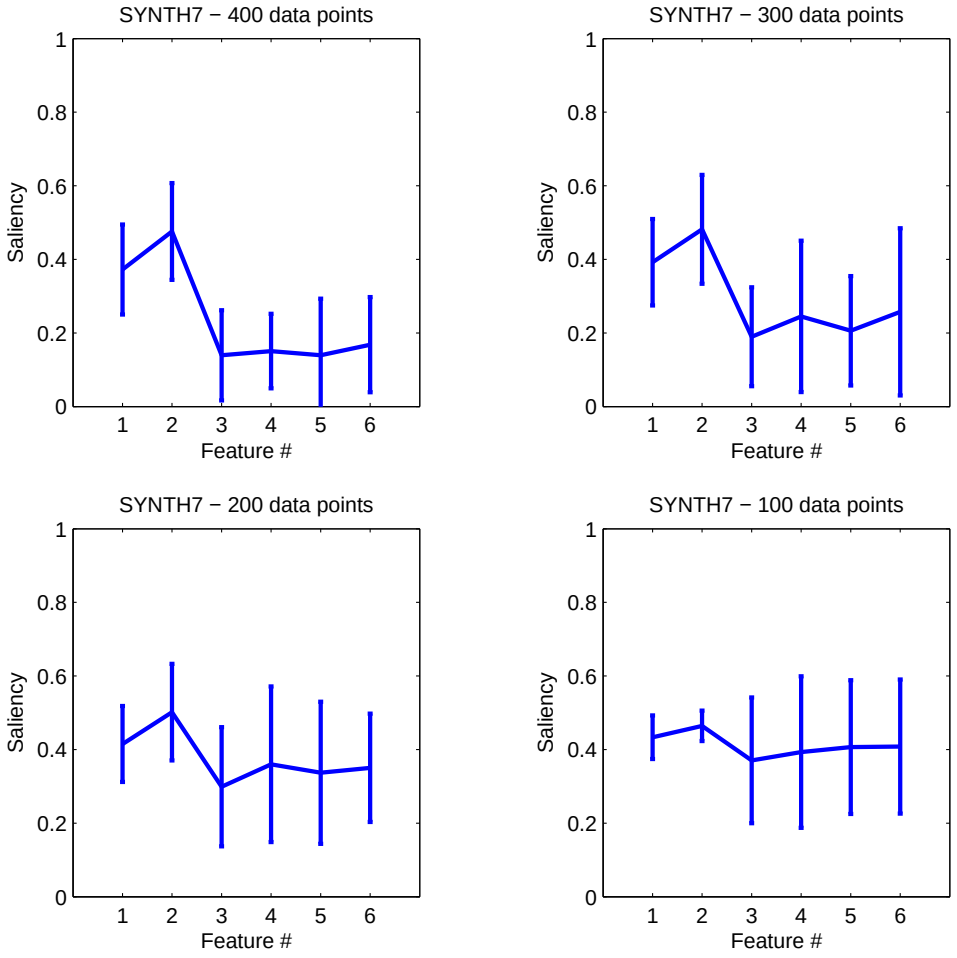

FIGURE A.10: Experimental results for different *SYNTH7* sample sizes (continuation of fig. A.9). Representation as in previous figures.

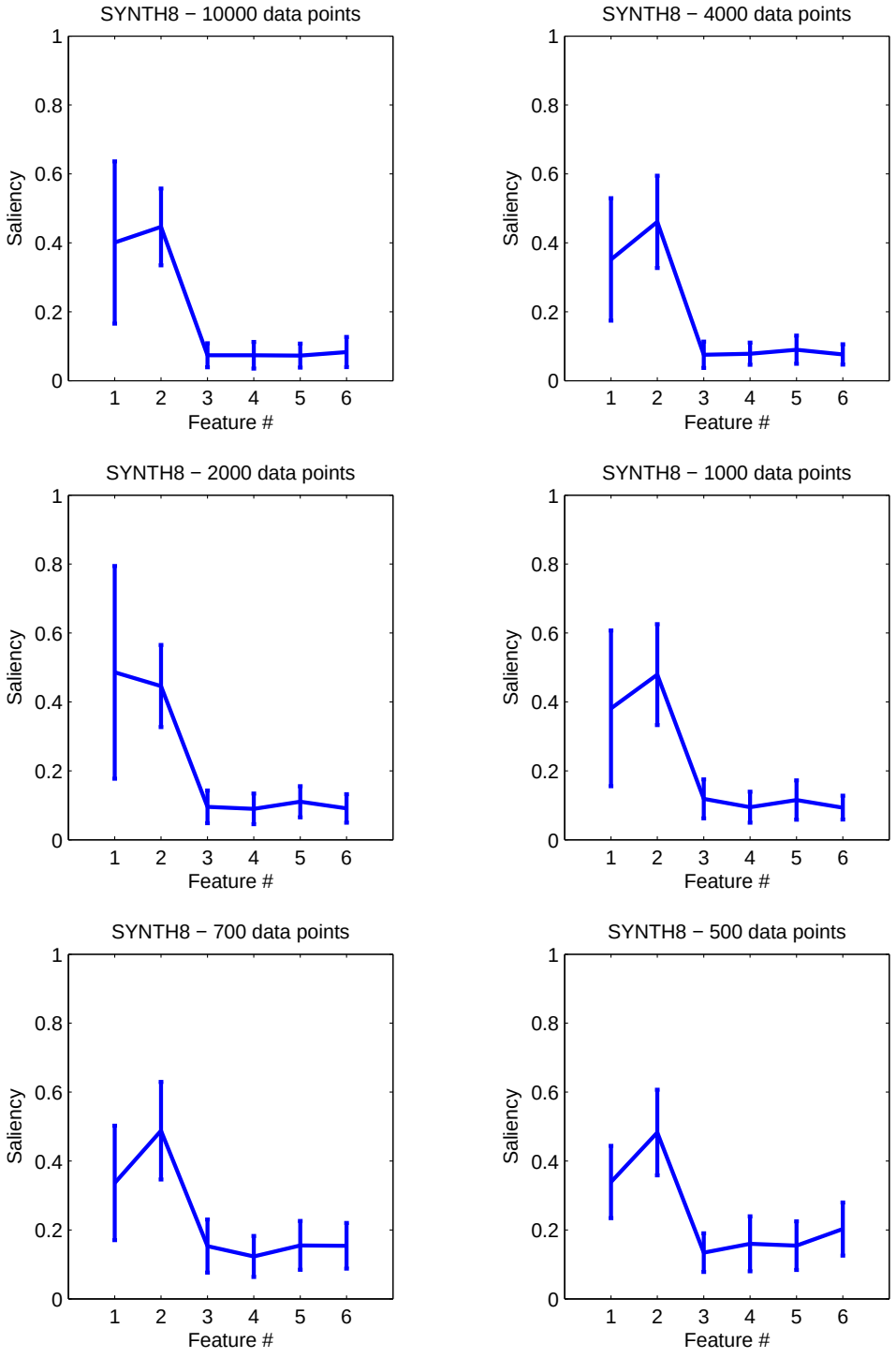

FIGURE A.11: Experimental results for different *SYNTH8* sample sizes (indicated in the plot titles). Representation as in previous figures.

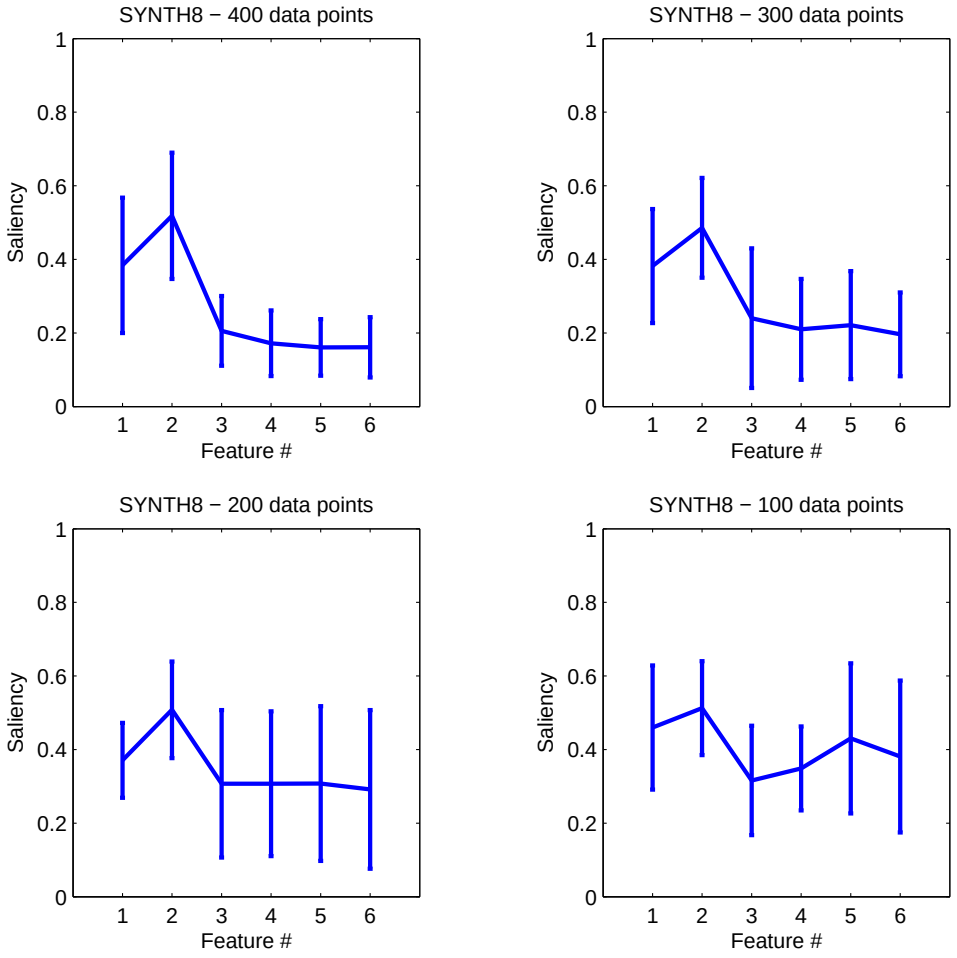

FIGURE A.12: Experimental results for different *SYNTH8* sample sizes (continuation of fig. A.11). Representation as in previous figures.

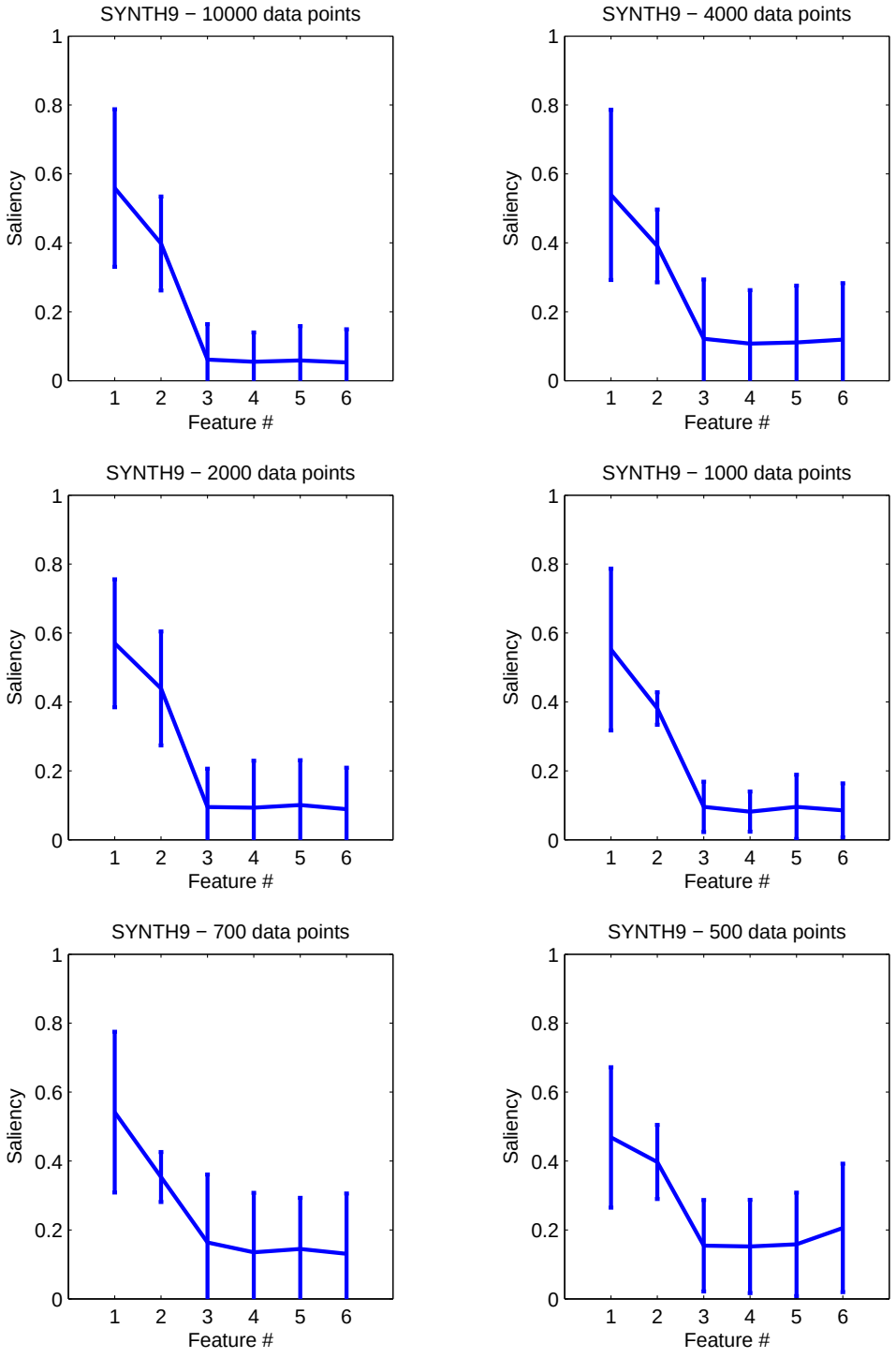

FIGURE A.13: Experimental results for different *SYNTH9* sample sizes (indicated in the plot titles). Representation as in previous figures.

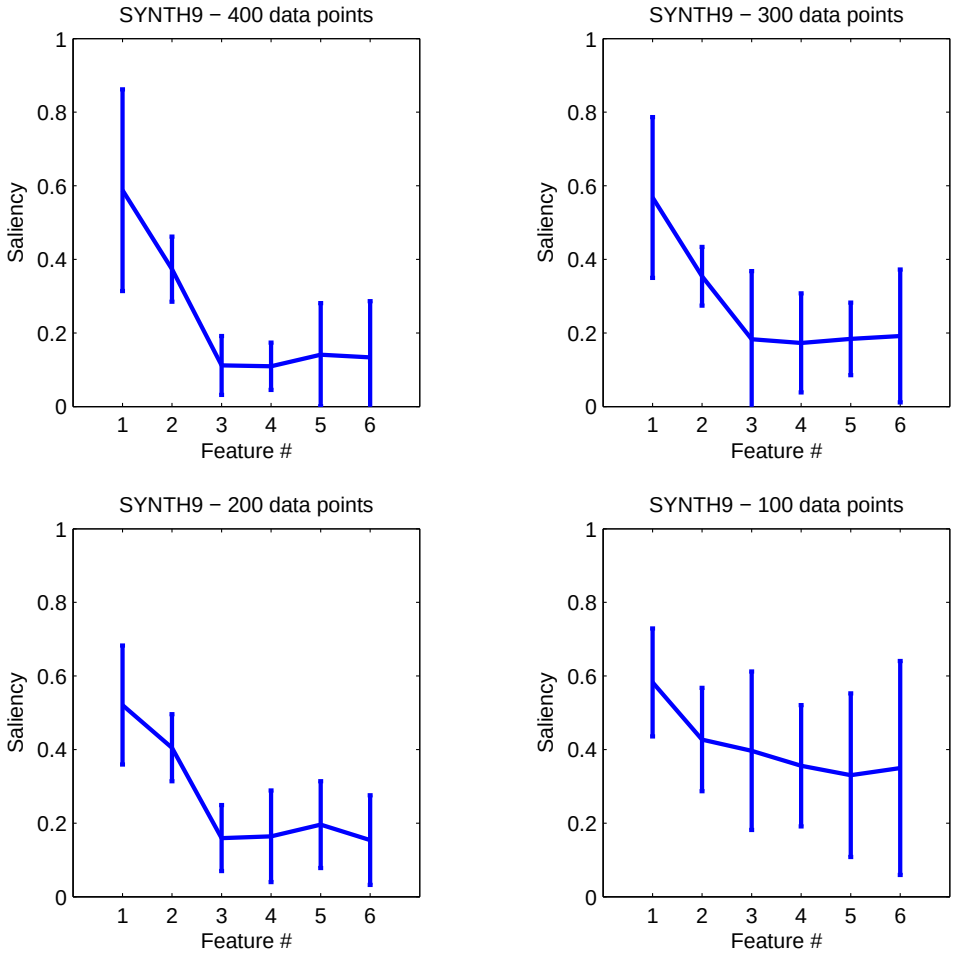

FIGURE A.14: Experimental results for different *SYNTH9* sample sizes (continuation of fig. A.13). Representation as in previous figures.

Appendix B

Figures: Noise effect

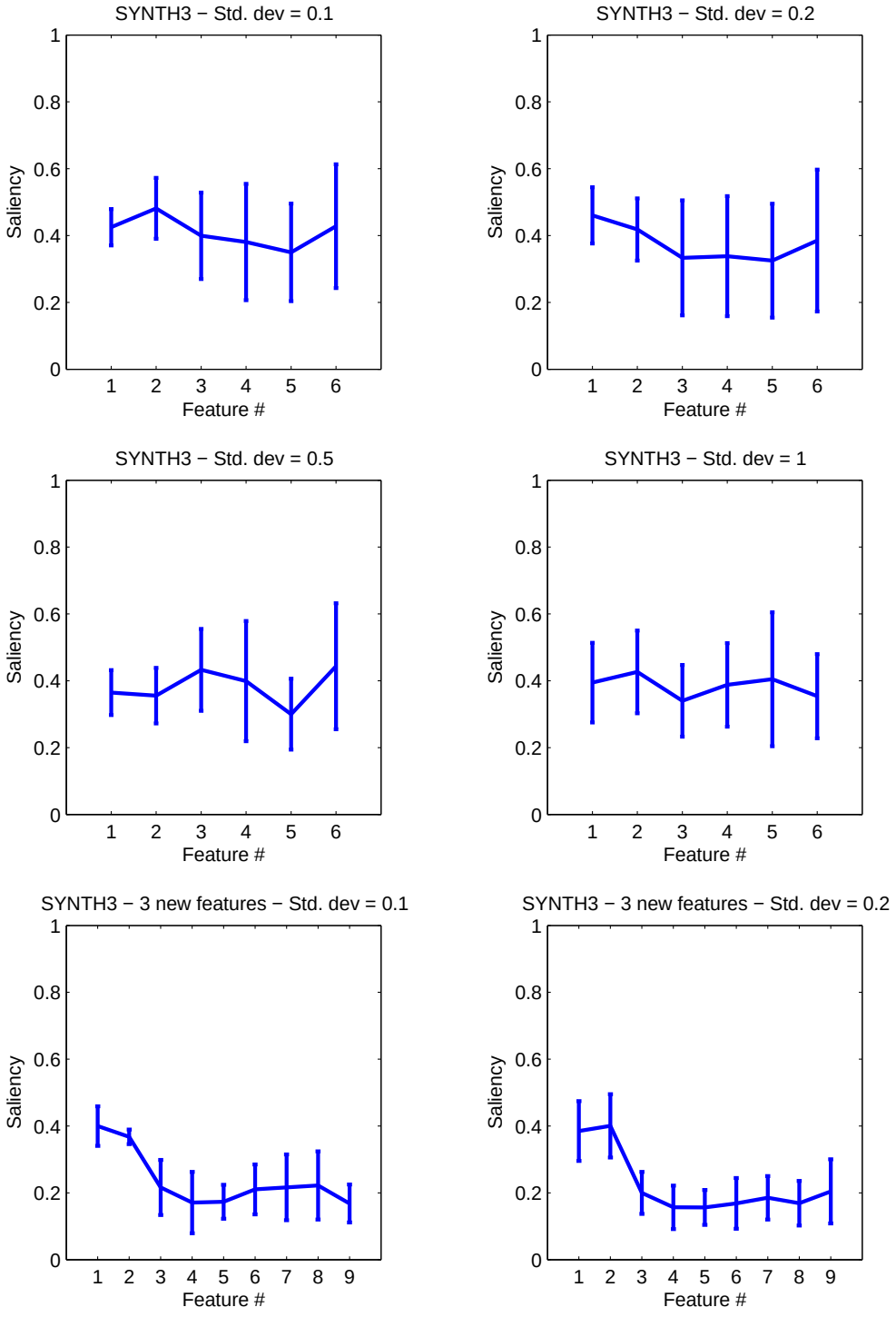

FIGURE B.1: Experimental results for a sample size of 1000 points from *SYNTH3*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

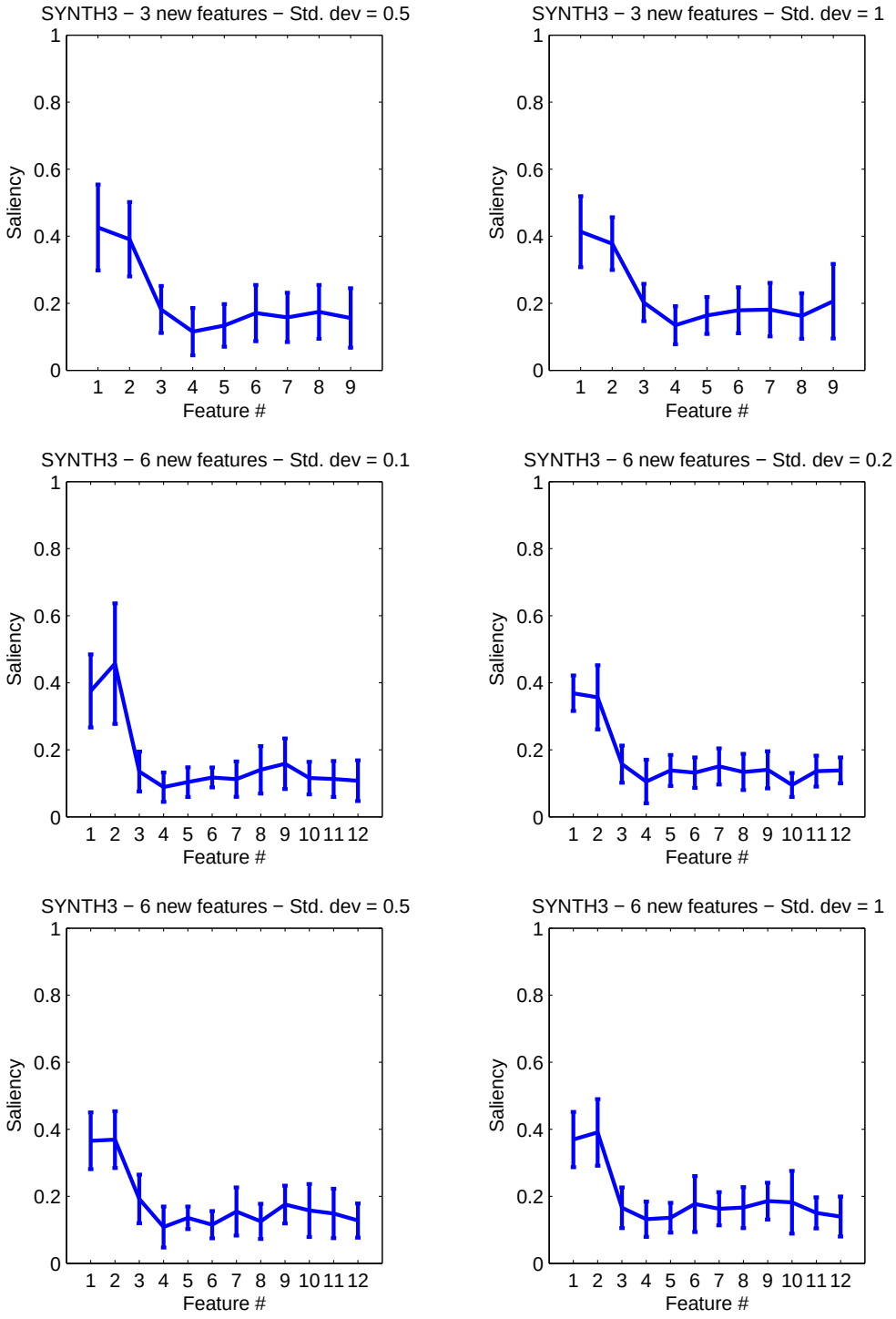

FIGURE B.2: Experimental results for a sample size of 1000 points from *SYNTH3*, to which different levels of Gaussian noise (continuation of fig. B.1) are added. Representation as in previous figures.

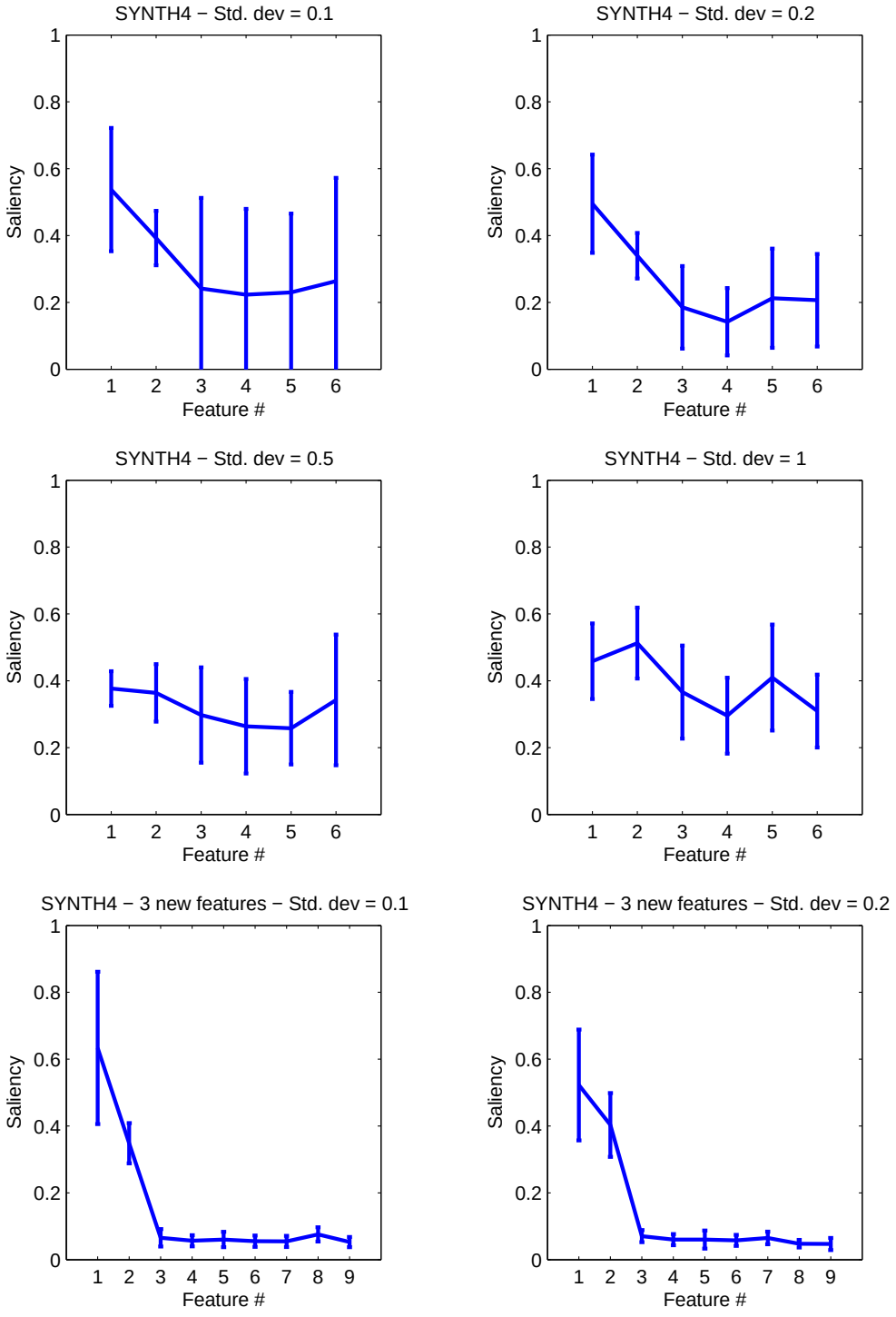

FIGURE B.3: Experimental results for a sample size of 1000 points from *SYNTH4*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

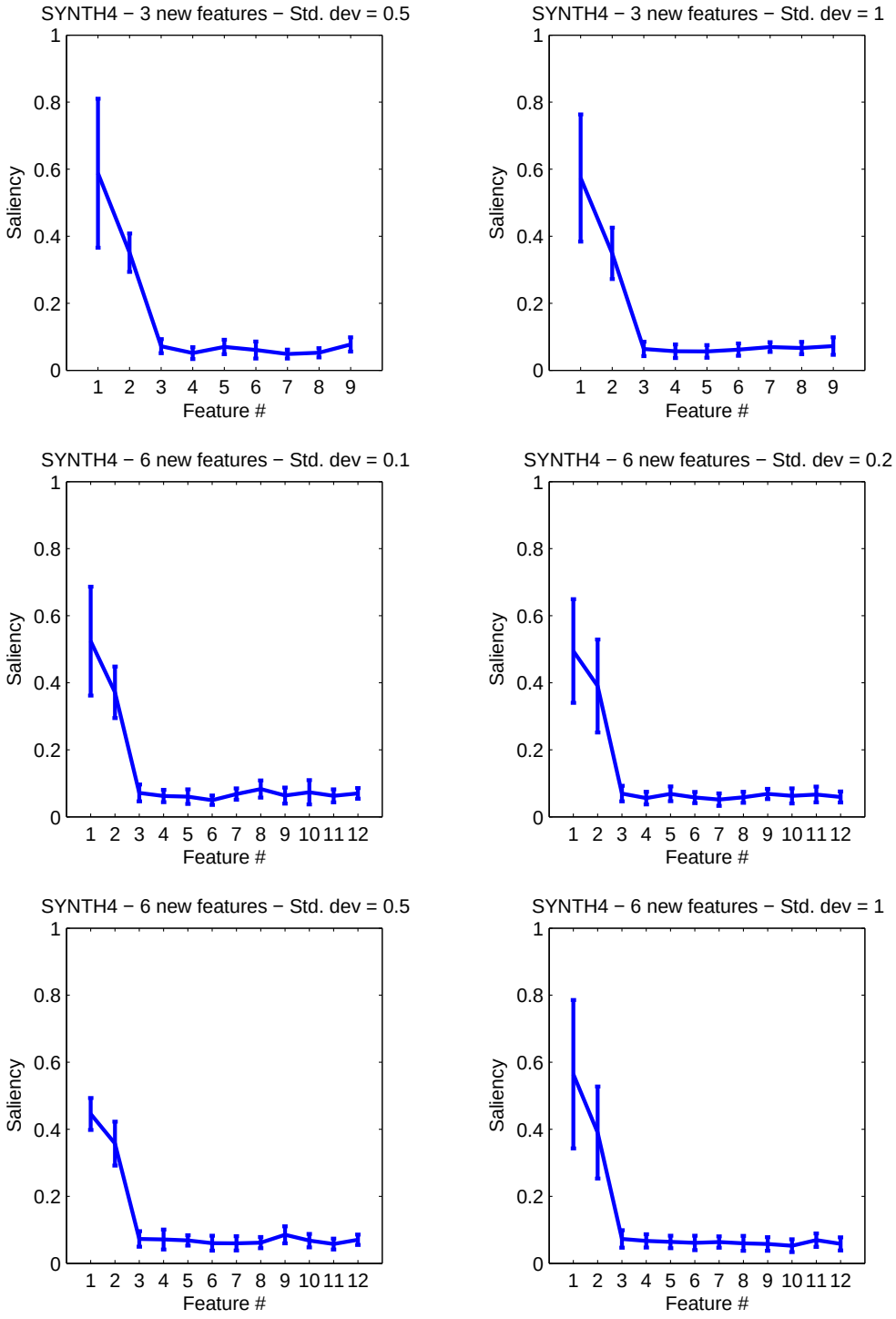

FIGURE B.4: Experimental results for a sample size of 1000 points from *SYNTH4*, to which different levels of Gaussian noise (continuation of fig. B.3) are added. Representation as in previous figures.

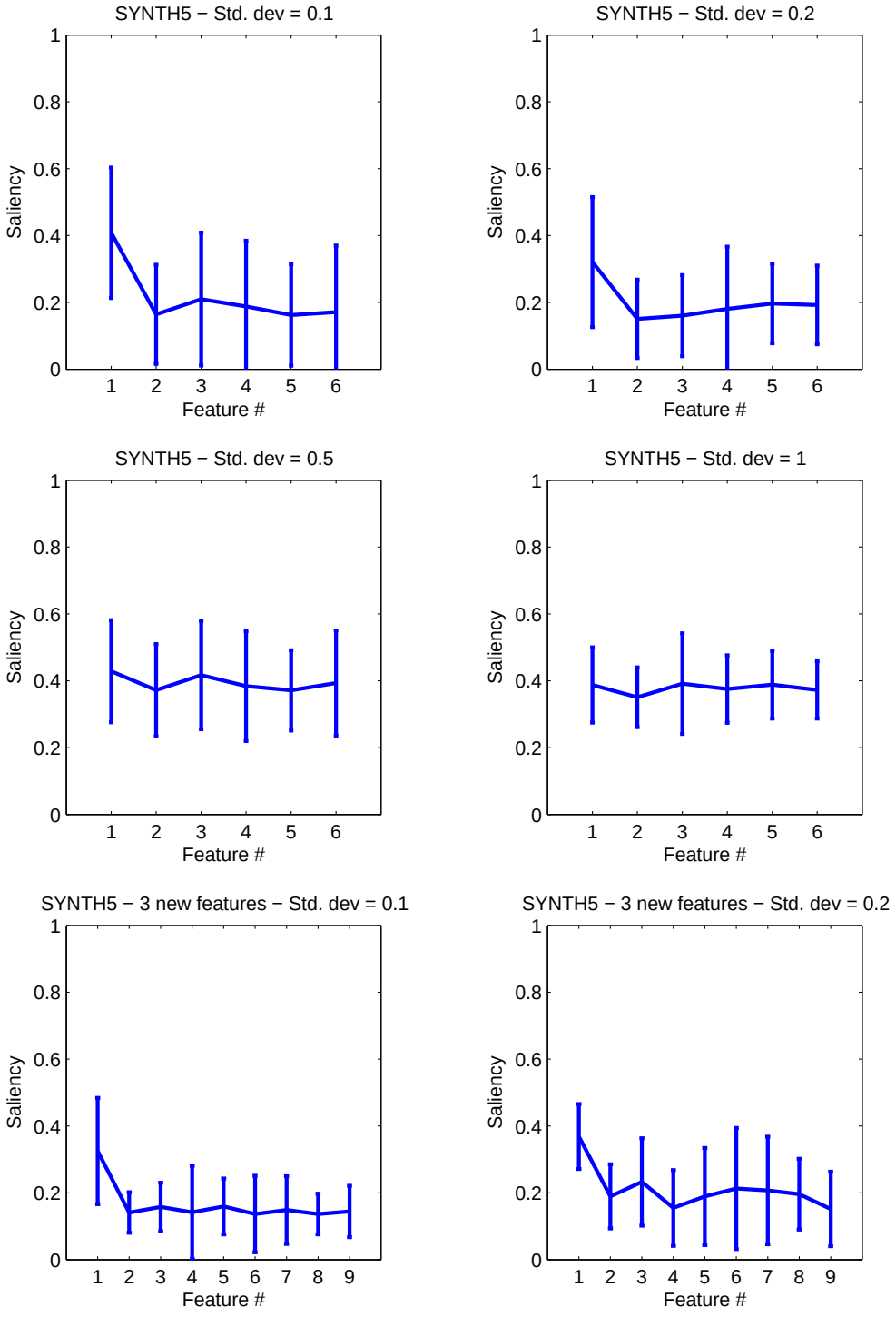

FIGURE B.5: Experimental results for a sample size of 1000 points from *SYNTH5*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

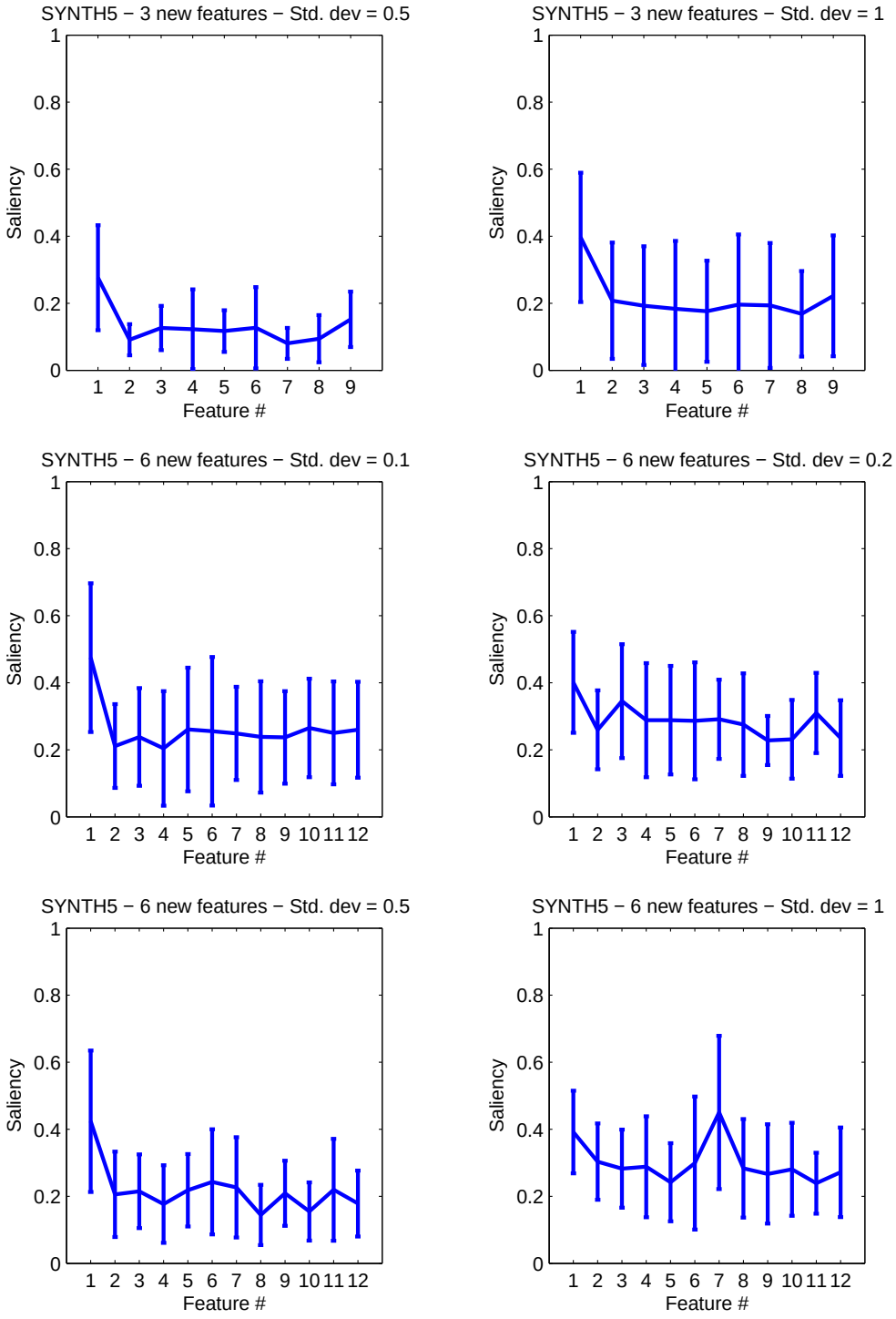

FIGURE B.6: Experimental results for a sample size of 1000 points from *SYNTH5*, to which different levels of Gaussian noise (continuation of fig. B.5) are added. Representation as in previous figures.

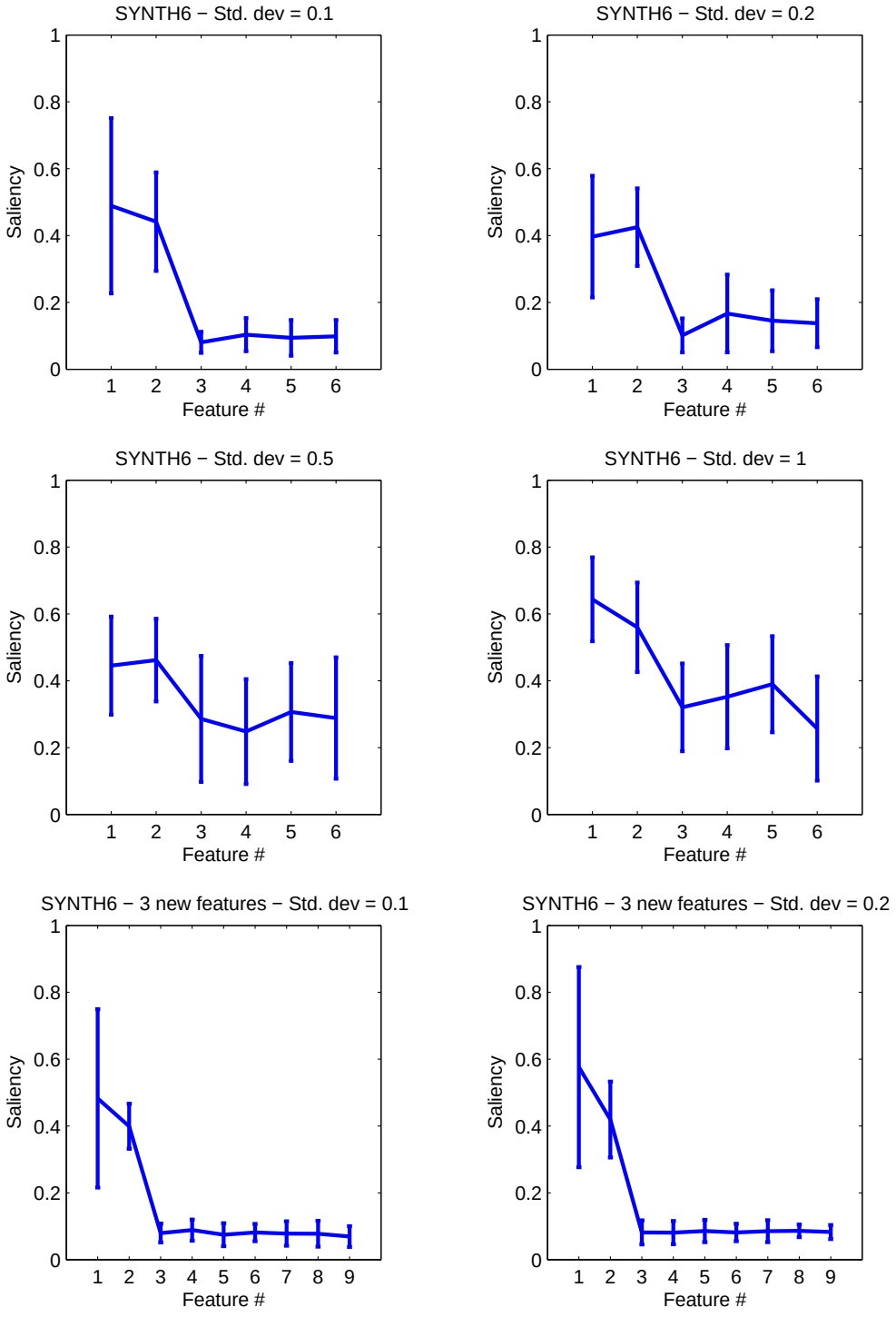

FIGURE B.7: Experimental results for a sample size of 1000 points from *SYNTH6*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

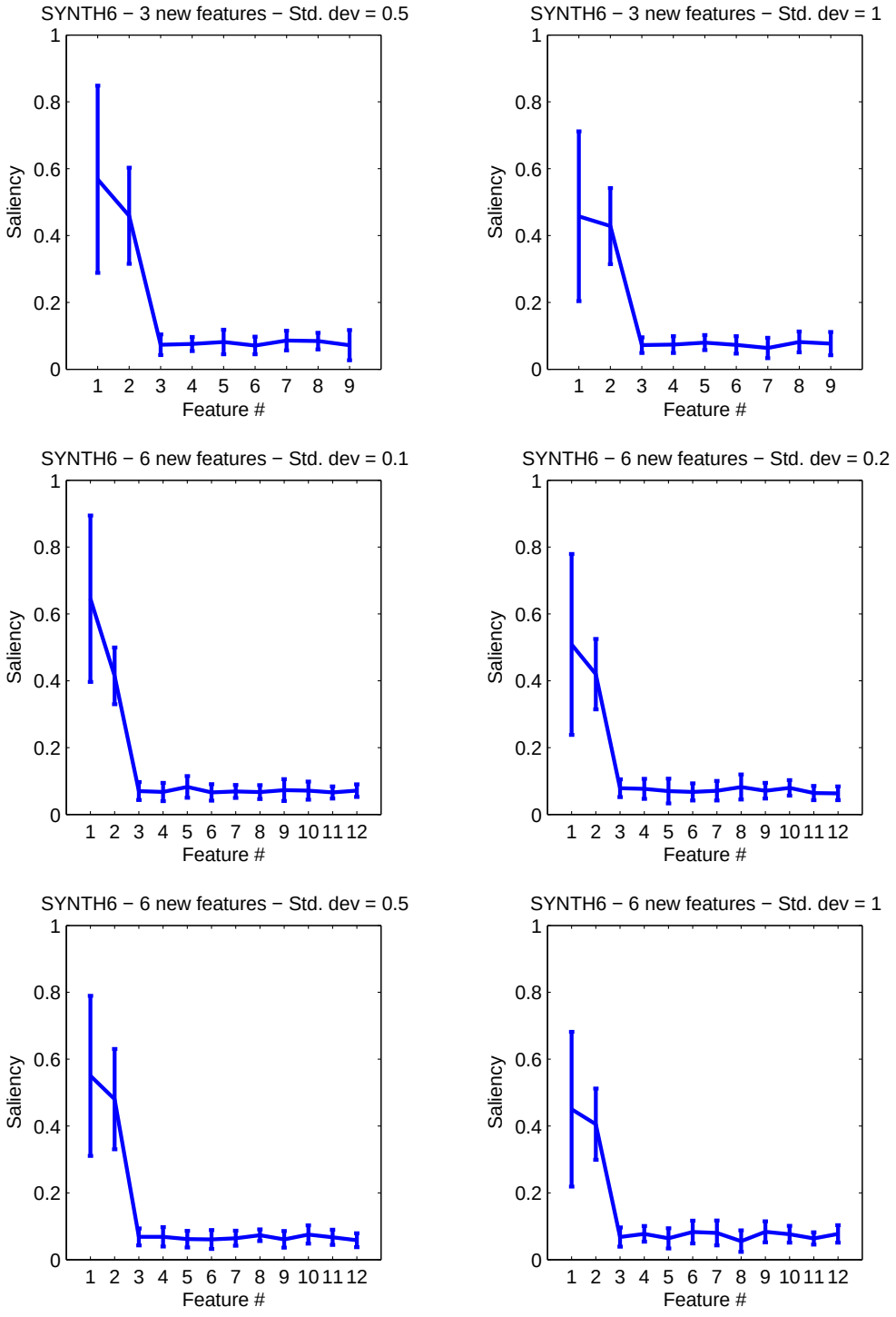

FIGURE B.8: Experimental results for a sample size of 1000 points from *SYNTH6*, to which different levels of Gaussian noise (continuation of fig. B.7) are added. Representation as in previous figures.

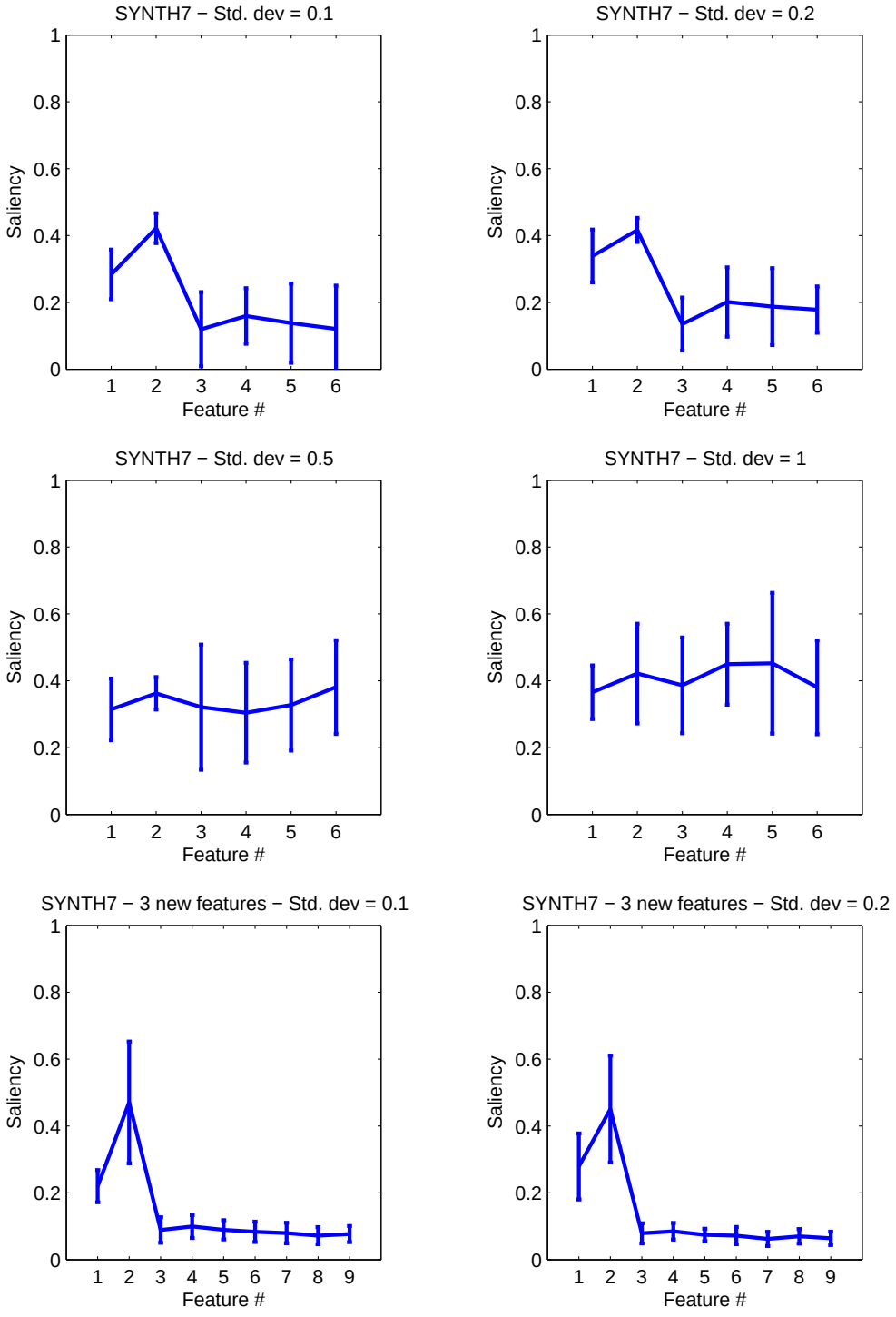

FIGURE B.9: Experimental results for a sample size of 1000 points from *SYNTH7*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

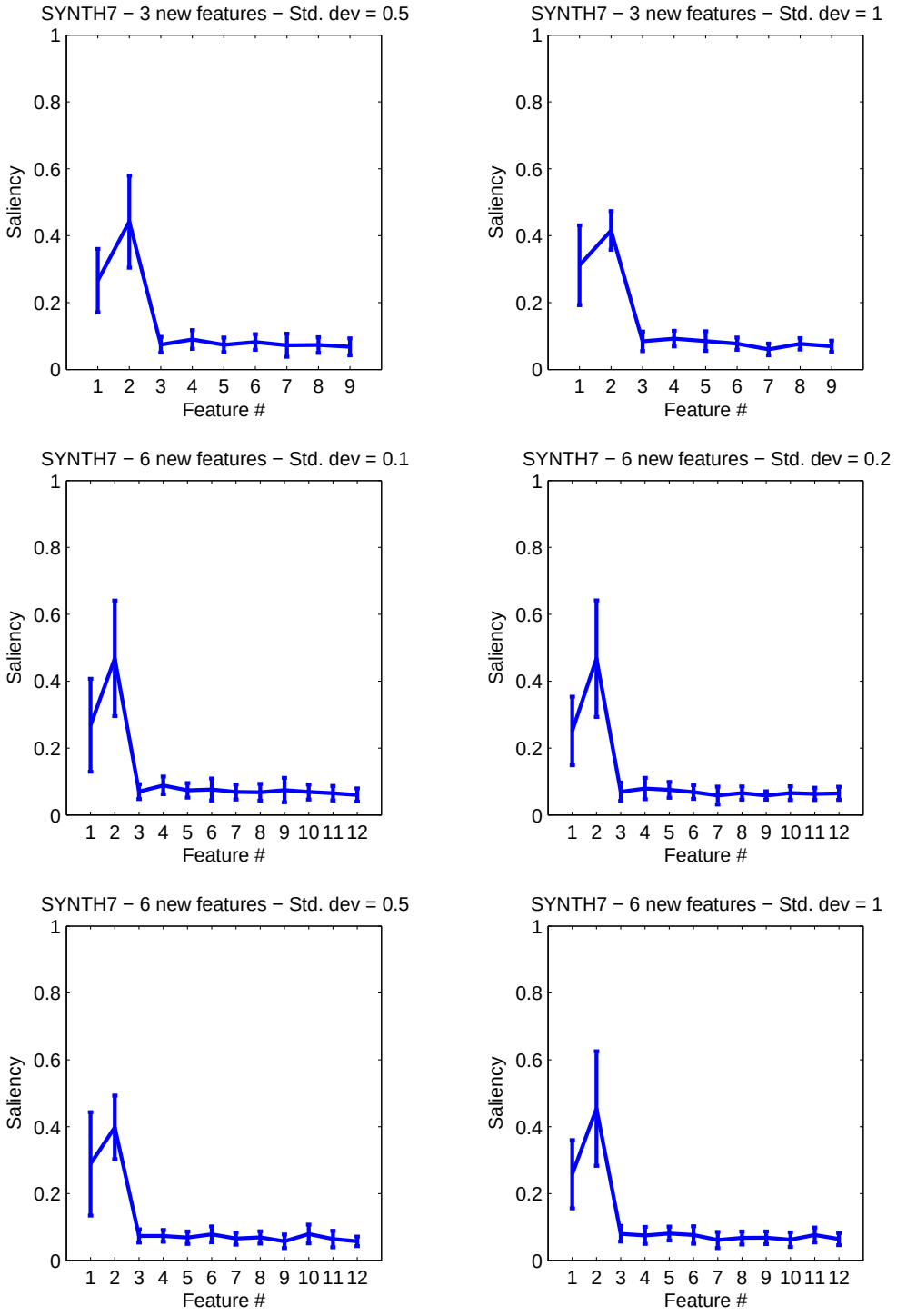

FIGURE B.10: Experimental results for a sample size of 1000 points from *SYNTH7*, to which different levels of Gaussian noise (continuation of fig. B.9) are added. Representation as in previous figures.

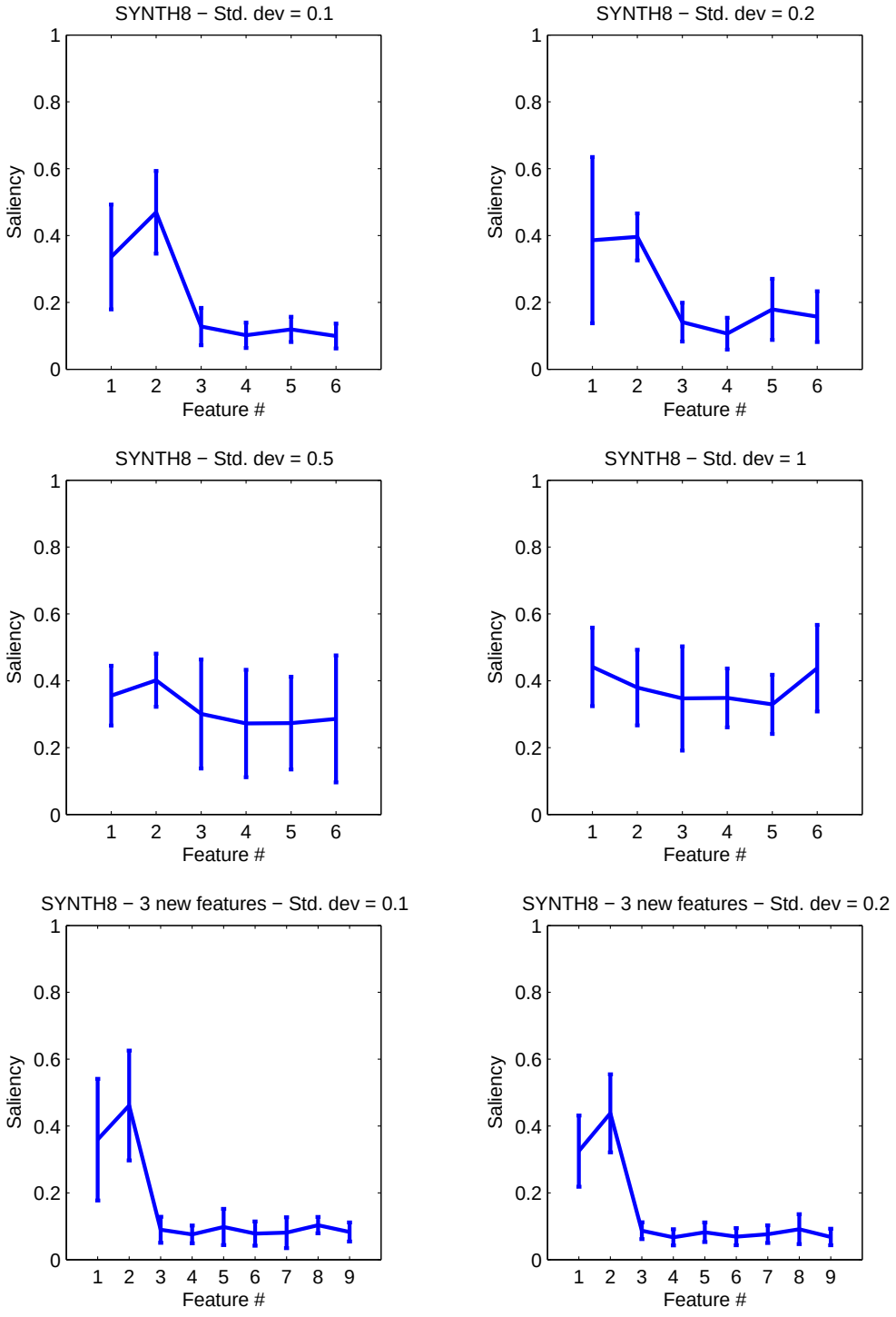

FIGURE B.11: Experimental results for a sample size of 1000 points from *SYNTH8*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

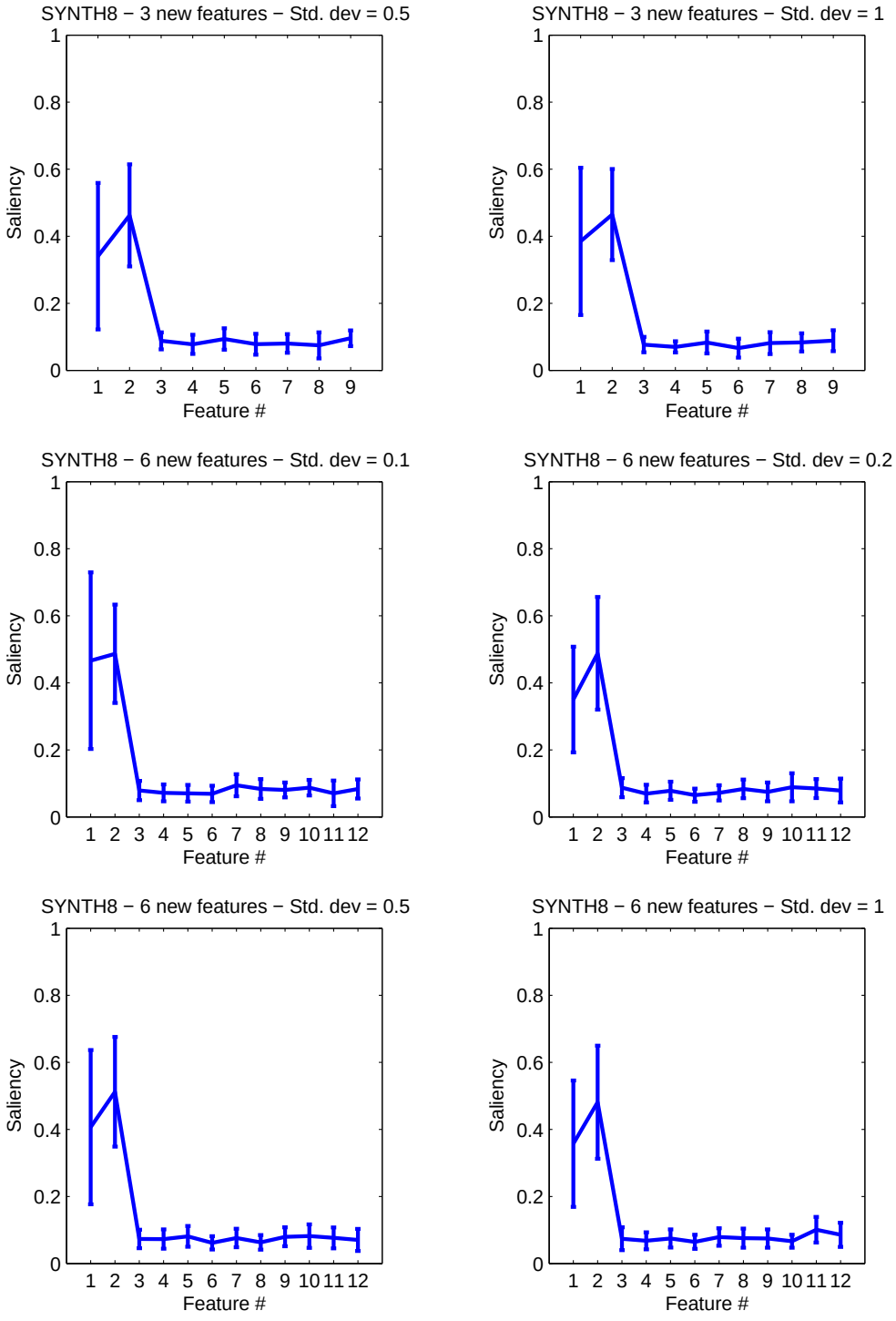

FIGURE B.12: Experimental results for a sample size of 1000 points from *SYNTH8*, to which different levels of Gaussian noise (continuation of fig. B.11) are added. Representation as in previous figures.

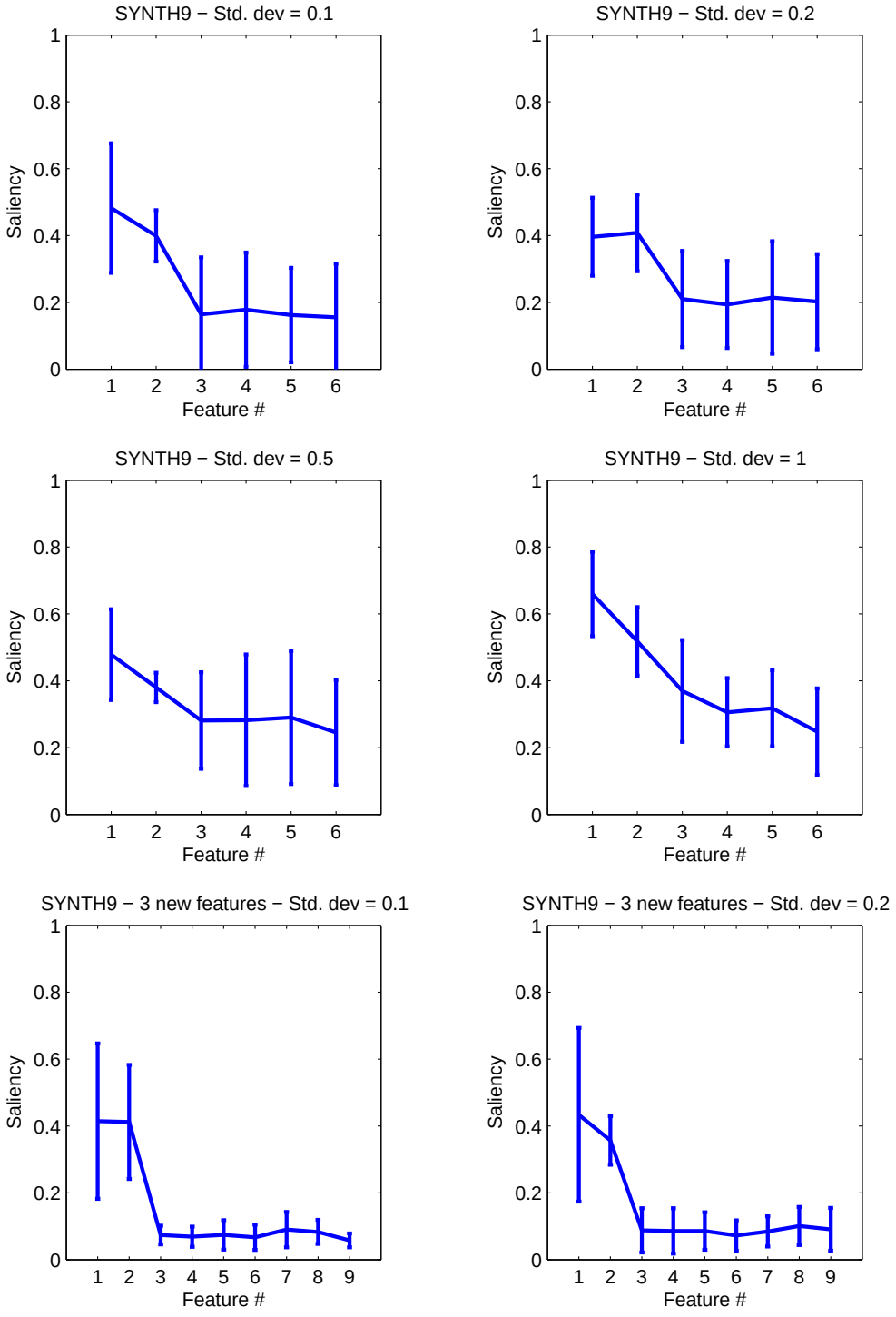

FIGURE B.13: Experimental results for a sample size of 1000 points from *SYNTH9*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

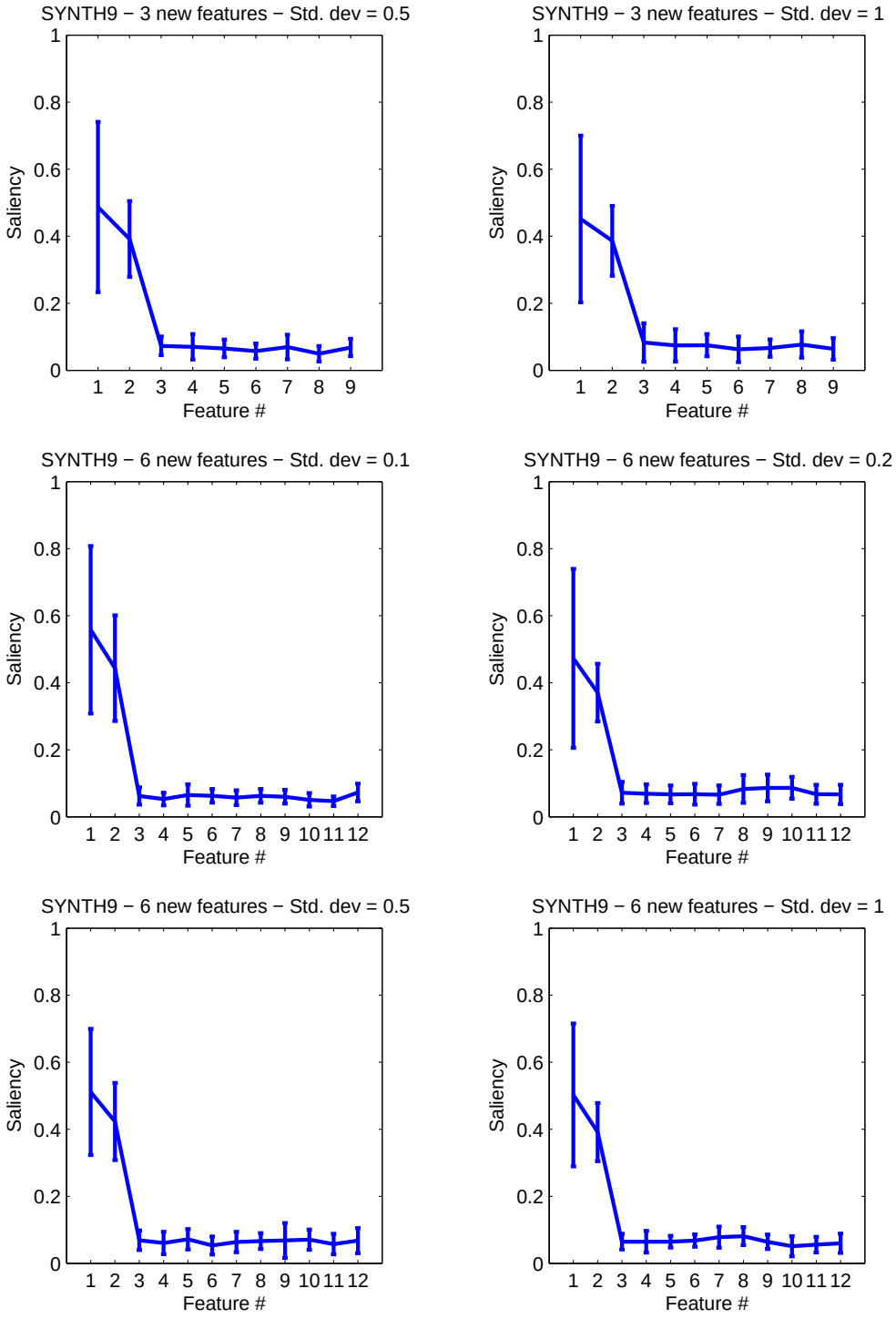

FIGURE B.14: Experimental results for a sample size of 1000 points from *SYNTH9*, to which different levels of Gaussian noise (continuation of fig. B.13) are added. Representation as in previous figures.

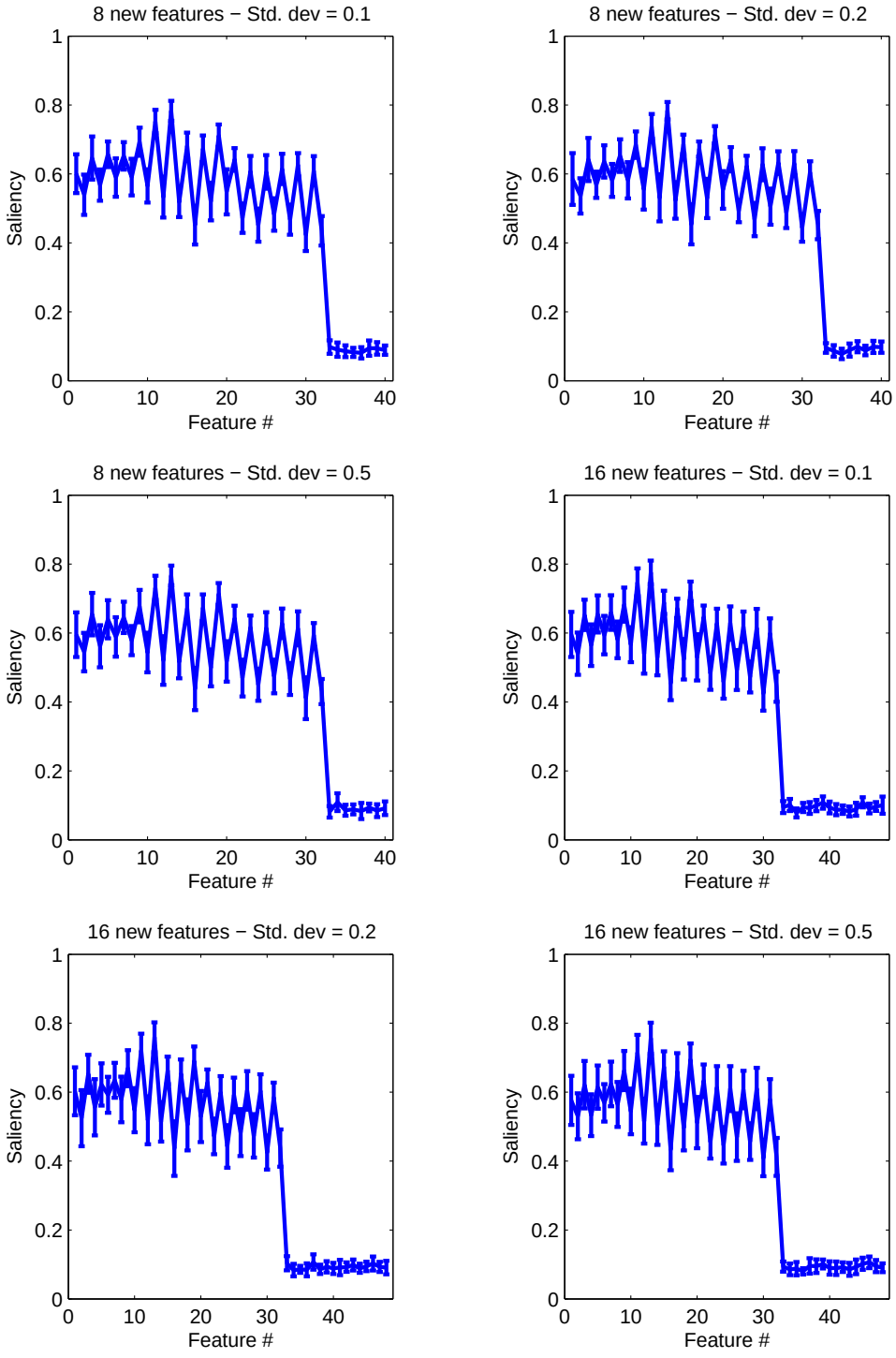

FIGURE B.15: Experimental results for a sample size of 351 points from *IONOSPHERE*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

Appendix C

Publications

- Vellido, A., Velazco, J. Assessment of the effect of noise on an unsupervised feature selection method for Generative Topographic Mapping. In 10th International Conference on Enterprise Information Systems (ICEIS 2008). Accepted for publication.
- Vellido, A., Velazco, J. The effect of noise and sample size on an unsupervised feature selection method for manifold learning. In Procs. of the International Joint Conference on Neural Networks (IJCNN 2008). LNCS.

The Effect of Noise and Sample Size on an Unsupervised Feature Selection Method for Manifold Learning

Alfredo Vellido and Jorge S. Velazco

***Abstract*— The research on unsupervised feature selection is scarce in comparison to that for supervised models, despite the fact that this is an important issue for many clustering problems. An unsupervised feature selection method for general Finite Mixture Models was recently proposed and subsequently extended to Generative Topographic Mapping (GTM), a manifold learning constrained mixture model that provides data visualization. Some of the results of a previous partial assessment of this unsupervised feature selection method for GTM suggested that its performance may be affected by insufficient sample size and by noisy data. In this brief study, we test in some detail such limitations of the method.**

I. Introduction

THE fields of machine learning and statistics coexist with data analysis as a common target and they overlap in what has come to be defined as Statistical Machine Learning. An example of this can be found in Finite Mixture Models, which are flexible and robust methods for multivariate data clustering [1]. The addition of visualization capabilities would benefit these models in many application scenarios, helping to provide intuitive cues about data structural patterns. One way to endow Finite Mixture Models with data visualization is by constraining the mixture components to be centered in a low-dimensional manifold embedded into the multivariate data space, as in Generative Topographic Mapping (GTM) [2]. This is a manifold learning model for simultaneous data clustering and visualization.

The interpretability of the clustering results provided by GTM becomes difficult when the analyzed data sets consist of a large number of features. This limitation can be overcome with methods to estimate the ranking of the data features according to their relative relevance, leading to feature selection (FS). The research on unsupervised FS is scarce in comparison to that for supervised models, despite the fact that FS becomes an issue of paramount importance for many clustering problems, regardless the unavailability of class labels. The interpretability of the clusters obtained by unsupervised methods would be improved by their description in terms of a reduced subset of relevant variables.

An important advance on unsupervised FS for Finite Mixture Models was presented in [3] and recently extended to GTM in [4] and to one of its variants for time series analysis in [5]. This method was preliminarily assessed in [6], where some of the results suggested that the performance of the method may be degraded by characteristics of the data such as insufficient sample size and the presence of noise. In this brief study, we provide far more detailed evidence of the limitations of the method through controlled experiments using synthetic data.

The remaining of the paper is organized as follows. First, brief introductions to the standard Gaussian GTM and its extension for Feature Relevance Determination (FRD) are provided in section 2. This is followed, in section 3, by a description of the experimental settings and, in section 4, by a presentation and discussion of the results. The paper closes with a brief summary of conclusions.

Department of Computing Languages and Systems (LSI). Technical University of Catalonia (UPC). C. Jordi Girona, 1-3. 08034, Barcelona, Spain (email: {avellido, e00728496}@lsi.upc.edu).

Alfredo Vellido is a researcher within the Ramón y Cajal program of the Spanish Ministry of Education and Science (MEC) and acknowledges funding from the MEC I+D project TIN2006-08114

II. Feature Relevance Determination for GTM

A. The Standard GTM Model

The neural network-inspired GTM is a manifold learning model with sound foundations in probability theory. It performs simultaneous clustering and visualization of the observed data through a nonlinear and topology-preserving mapping from a visualization latent space in $\Re^L$ (with *L* being usually 1 or 2 for visualization purposes) onto a manifold embedded in the $\Re^D$ space, where the observed data reside.

For each feature *d*, the functional form of this mapping is the generalized linear regression model $\mathrm{y}_d(\mathbf{u}, \mathbf{W}) = \sum_m^M \phi_m(\mathbf{u}) w_{md}$, where ϕ_m is one of *M* basis functions, defined here as spherically symmetric Gaussians, generating the non-linear mapping from a latent vector $\mathbf{u}$ to the manifold in $\Re^D$. The matrix $\mathbf{W}$ of adaptive weights w_{md} explicitely defines this mapping.

The prior distribution of $\mathbf{u}$ in latent space is constrained to form a uniform discrete grid of *K* centres. A density model in data space is therefore generated for each component *k* of the mixture, which, assuming that the observed data set $\mathbf{X}$ is constituted by *N* independent, identically distributed (i.i.d.) data points $\mathbf{x}_n$, leads to the definition of a complete log-likelihood in the form:

$$L(\mathbf{W},\beta|\mathbf{X}) = \sum_{n=1}^{N} \ln\Big\{\frac{1}{K}\sum_{k=1}^{K}\Big(\frac{\beta}{2\pi}\Big)^{D/2}\exp\Big\{-\beta/2\|\mathbf{y}_k-\mathbf{x}_n\|^2\Big\}\Big\} \quad (1)$$

where $\mathbf{y}_k$ is a reference or prototype vector consisting of elements ($y_{dk} = \sum_m^M \phi_m(\mathbf{u}_k) w_{md}$), which are an instantiation of the generalized linear regression model described above. From Eq. (1), the adaptive parameters of the model, which are $\mathbf{W}$ and the common inverse variance of the Gaussian components, β, can be optimized by maximum likelihood

(ML) using the Expectation-Maximization (EM) algorithm. Details can be found in [2].

B. The FRD-GTM

The problems of feature selection and feature relevance determination are commonly understood as one of the possible strategies for data dimensionality reduction, usually for supervised problems. In such setting, a data feature is said to be relevant (and it is eventually selected) only if its absence (or its absence in combination with the absence of others) worsens significantly the classification or predictive performance of the defined model. Feature selection and feature relevance determination for unsupervised learning, even if sharing the dimensionality reduction objective of their supervised counterparts, are far less investigated problems. Here, the relevance is not longer related to a label or target variable, and various feature ranking criteria can be considered, including, but not limited to, *saliency, entropy, smoothness, density* and reliability [7].

In this paper, unsupervised feature relevance is understood as the likelihood of a feature being responsible for generating the data cluster structure. Therefore, relevant features will be those which better separate the natural clusters in which the data are structured. Moreover, we are interested in unsupervised feature selection methods that are suitable for clustering models that also provide data visualization. With that in mind, the FRD technique was defined for the GTM model in [4]. For the unsupervised GTM clustering model, relevance is defined through the concept of saliency.

The FRD problem was investigated for GTM in [4]. Feature relevance in this unsupervised setting is understood as the likelihood of a feature being responsible for generating the data cluster structure. In this unsupervised setting, relevance is defined through the concept of saliency. Formally, the saliency of feature *d* can be defined as $\rho_d = P(\eta_d = 1)$, where η=$(\eta_1, \ldots, \eta_D)$ is a set of binary indicators that can be integrated in the EM algorithm as missing variables. A value of $\eta_d = 1$ $(\rho_d = 1)$ indicates that feature *d* has the maximum possible relevance. According to this definition, the FRD-GTM mixture density can be written as:

$$p(\mathbf{x}|\mathbf{W},\beta,\mathbf{w}_0,\boldsymbol{\beta}_0,\boldsymbol{\rho}) = \sum_{k=1}^{K} \frac{1}{K} \prod_{d=1}^{D} \{\rho_d p(x_d|\mathbf{u}_k;\mathbf{w}_d,\beta) + (1-\rho_d) q(x_d|\mathbf{u}_0;w_{0,d},\beta_{0,d})\} \quad (2)$$

where $\mathbf{w}_d$ is the vector of $\mathbf{W}$ corresponding to feature *d* and $\boldsymbol{\rho} \equiv \{\rho_1, \ldots, \rho_D\}$. A feature *d* will be considered irrelevant, with *irrelevance* $(1 - \rho_d)$, if $p(x_d|\mathbf{u}_k; \mathbf{w}_d, \beta) = q(x_d|\mathbf{u}_0; w_{0,d}, \beta_{0,d})$ for all the mixture components *k*, where *q* is a common density followed by feature *d*. Notice that this is like saying that the distribution for feature *d* does not follow the cluster structure defined by the model. This common component requires the definition of two extra adaptive parameters in (2): $\mathbf{w}_0 \equiv \{w_{0,1}, \ldots, w_{0,D}\}$ and $\boldsymbol{\beta}_0 \equiv \{\beta_{0,1}, \ldots, \beta_{0,D}\}$ (so that $\mathbf{y}_0 = \phi_0(\mathbf{u}_0)\mathbf{w}_0$). For fully relevant ($\rho_d \to 1$) features, the common component variance vanishes:$(\beta_{0,d})^{-1} \to 0$. The parameters of the model can, once again, be optimized by ML using the EM algorithm. Detailed calculations can be found in [8].

III. Experimental Settings

The results of statistically principled models for probability density estimation, such as GTM and its variants, are bound to be affected, in one way or another, by sample size and by the presence of uninformative noise in the data. Here, we assess such effects on the FRD-GTM model described in the previous section. For that, data with very specific characteristics are required. We use synthetic sets similar to those in [3] for comparative purposes.

The first synthetic set (hereafter referred to as *synth1*) is a variation on the *Trunk* data set used in [3]), and was designed for its 10 features to be in decreasing order of relevance. It consists of data sampled from two Gaussians $N(\mu_1, \mathbf{I})$ and $N(\mu_2, \mathbf{I})$, where: $\left(\mu_1 = 1, \frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{2d-1}}, \ldots, \frac{1}{\sqrt{19}}\right)$ and $\mu_1 = -\mu_2$. We hypothesize (*H1*) that the feature relevance ranking estimated by FRD-GTM for these data will deteriorate gradually as sample size decreases. Samples of *synth1* of different sizes, from 100 to 10,000 points, were used in this study to test *H1*. It is also hypothesized (*H2*) that the feature relevance ranking will deteriorate in proportion to the level of noise. In order to test *H2*, four increasing levels of Gaussian noise, of standard deviations 0.1, 0.2, 0.5, and 1, were added to the 10 original features of *synth1*, for a given sample size.

The second dataset (hereafter referred to as *synth2*) consists of a contrasting combination of features: the first two define four neatly separated Gaussian clusters with centres located at $(0, 3), (1, 9), (6, 4)$ and $(7, 10)$; they are meant to be relatively relevant. The next four features are Gaussian noise and, therefore, rather irrelevant in terms of defining cluster structure. Similar experiments to the ones devised for *synth1* were designed to further test *H1* and *H2*.

The FRD-GTM parameters $\mathbf{W}$ and $\mathbf{w}_0$ were initialized with small random values sampled from a normal distribution. Saliencies were initialized at $\rho_d = 0.5, \forall d, d = 1, \ldots, D$. The grid of GTM latent centres was fixed to a square layout of 3×3 nodes (i.e., 9 constrained mixture components). The corresponding grid of basis functions ϕ_m was fixed to a 2×2 layout.

IV. Experimental results and discussion

The experiments outlined in the previous section aim to assess the effect of sample size and the presence of noise on the performance of FRD-GTM in the process of unsupervised feature relevance estimation.

A. The Effect of Sample Size

The FRD ranking results for *synth1* are shown in Fig. 1, for sample sizes from 10,000 down to 100 points. Further sample sizes were tested, conforming to a similar pattern; their results are not included for the sake of brevity. A deterioration of the results is clearly observed for datasets of less than 1,000 points.

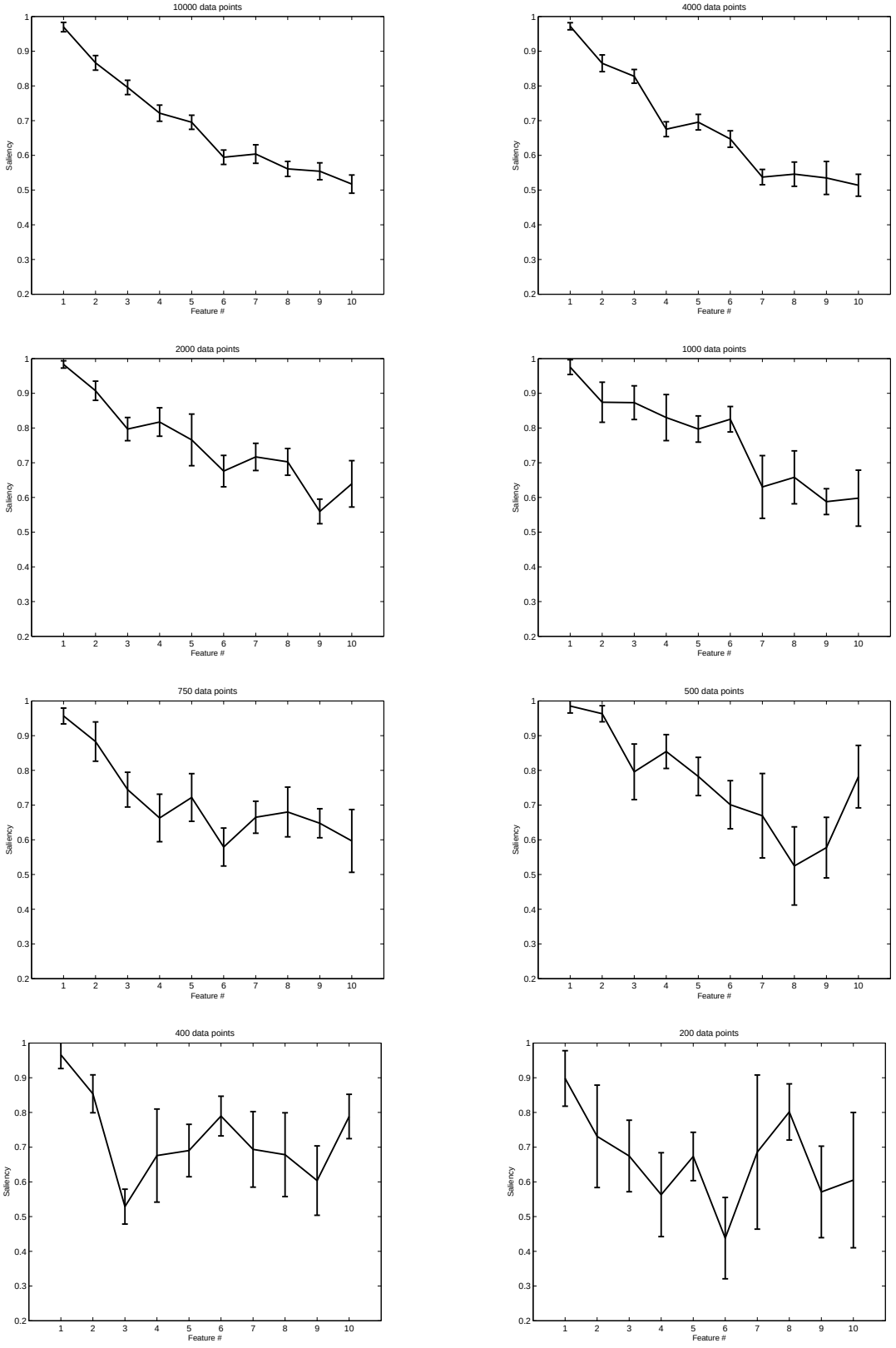

Fig. 1. Experiments with different *synth1* sample sizes (indicated in the plot titles) Mean saliencies ρ_d for the 10 features. The bars span from the mean minus to the mean plus one standard deviation of the saliencies over 20 runs of the algorithm.

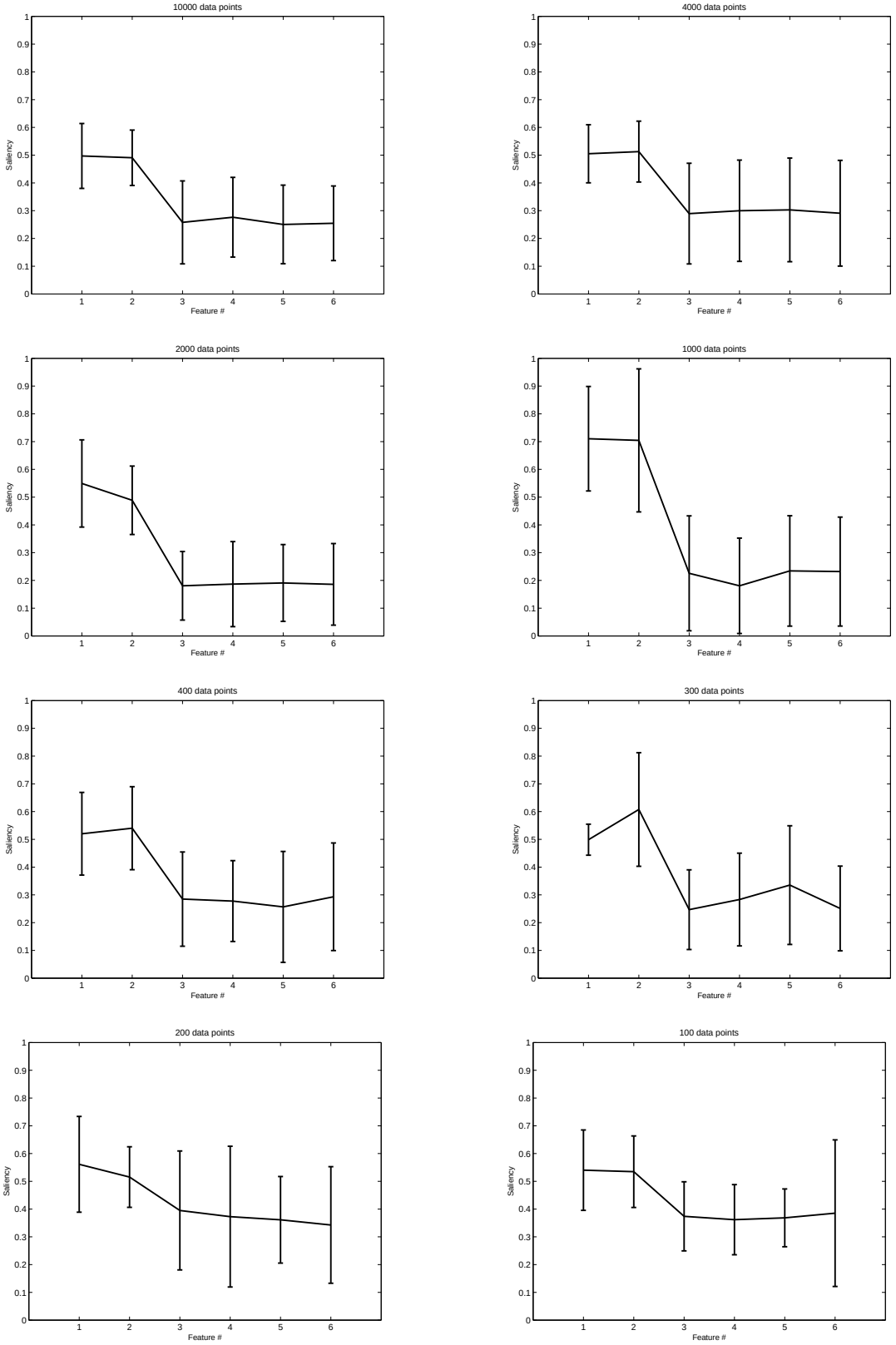

Fig. 2. Experimental results for different *synth2* sample sizes (indicated in the plot titles). Representation as in previous figures.

This deterioration takes two forms: Firstly, a breach of the expected monotonic decrease of the mean feature saliencies. Secondly, a neat increase of uncertainty in the results, illustrated in Fig. 1 in the form of bigger bars of the standard deviation of the estimated saliencies. As a result, the confidence on the validity of the results for small sample sizes decreases considerably. According to these results, *H1* is at least partially supported.

The FRD ranking results for *synth2*, again for sample sizes from 10.000 down to 100 points, are shown in Fig. 2. This is an easier problem for the model, and this is reflected by the fact that the saliency estimated for the two first features is higher than that estimated for the rest of the features, even for a sample size as small as 100 points. A deterioration of the saliency estimation is nevertheless evident for the smallest of the sample sizes investigated. This is consistent with the results for *synth1* and, again, *H1* is partially supported.

B. The effect of Noise

In the experiments reported in Fig. 3, four levels of Gaussian noise of increasing level were added to a sample of 1,000 points of *synth1*. The FRD-GTM is shown to behave robustly even in the presence of a substantial amount of noise, although its performance deteriorates significantly for noise of standard deviation = 1, as reflected in the breach of the expected monotonic decrease of the mean feature saliencies. *H2* is, therefore, partially supported by these results.

Fig. 4 displays the results of a similar experiment for *synth2*. They are fully consistent with those obtained with *synth1*. The model again behaves robustly in the presence of noise and clearly deteriorates at the highest level of added noise, for which the model struggles to distinguish the first two features from the purely noisy ones. Hypothesis *H2* is, again, at least partially supported.

This support for hypothesis *H2* is, even if partial, certainly not unexpected. As robust as it may be, the FRD-GTM model is still prone to data overfitting. That is, at some point, the model will start learning the noise as much as learning the underlying signal distributions. The resulting FRD-GTM model will be over-complex and, if the noise is uninformative (i.e., in this case, if the noise affects all data features equally), the method of relevance determination will eventually start struggling to provide correct saliency estimations. One way around this problem is to endow the model with regularization capabilities to effectively control complexity [9], [10], [11]. FRD-GTM is thus likely to benefit from the definition of extensions of the model encompassing adaptive regularization.

V. Conclusions

In this paper, the effects of sample size and the presence of noise on a method of unsupervised feature relevance determination for the manifold learning GTM model, have been investigated in some detail. The FRD-GTM has been shown to behave with reasonable robustness even at small sample sizes and in the presence of a fair amount of noise. Even though, performance deterioration has been observed at very small sample sizes and in the presence of high level of noise.

This relative weakness of the method in the presence of noise makes it convenient to consider possible strategies for model regularization and, therefore, future research will be devoted the design of methods for automatic and proactive model regularization to prevent or at least limit the negative effect of data overfitting on the FRD method for GTM. Some of such methods have already been designed for the standard GTM formulation [9], [10] and could be extended to FRD-GTM. Alternatively, regularization could be accomplished through a reformulation of the GTM within a variational Bayesian theoretical framework [11]. Again, this could be extended to accomodate FRD.

Future research should also extend the current experimental design to include a wider variety of artificial data sets of different characteristics, as well as to include comparisons with alternative unsupervised feature relevance determitation and feature selection techniques.

References

[1] G. J. McLachlan and D. G. Peel. *Finite mixture models*. New York: John Wiley-Sons, 2000.

[2] C. M. Bishop, M. Svensén and C. K. I. Williams. “GTM: The Generative Topographic Mapping”. *Neural Computation*, 10(1), pp. 215–234, 1998.

[3] M. H. C. Law, M.A.T. “Figueiredo and A. K. Jain, Simultaneous Feature Selection and Clustering Using Mixture Models”, *IEEE T. Pattern Anal, 26(9)*, pp. 1154–1166, 2004.

[4] A. Vellido, P. J. G. Lisboa and D. Vicente, “Robust Analysis of MRS Brain Tumour Data Using t-GTM”, *Neurocomputing*, 69(7–9), pp. 754–768, 2006.

[5] I. Olier and A. Vellido, “Time Series Relevance Determination through a topology-constrained Hidden Markov Model”, *In Proc. of the 7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2006)*, Burgos, Spain. LNCS 4224, pp. 40–47, 2006.

[6] A. Vellido, “Assessment of an Unsupervised Feature Selection Method for Generative Topographic Mapping”, *16th International Conference on Artificial Neural Networks (ICANN 2006)*, Athens, Greece. LNCS 4132, pp. 361–370, 2006.

[7] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection”, *Journal of Machine Learning Research*, 3 (7–8) pp. 1157–1182, 2003.

[8] A. Vellido, “Preliminary theoretical results on a feature relevance determination method for Generative Topographic Mapping”, *Technical Report LSI-05-13-R*, Universitat Politecnica de Catalunya, UPC, Barcelona, Spain, 2005.

[9] C. M. Bishop, M. Svensén and C. K. I. Williams. “Developments of the Generative Topographic Mapping”, *Neurocomputing*, 21(1–3), pp. 203–224, 1998.

[10] A. Vellido, W. El-Deredy, and P. J. G. Lisboa, “Selective smoothing of the Generative Topographic Mapping”, *IEEE T. Neural Network*, 14(4), pp. 847–852, 2003.

[11] I. Olier and A. Vellido, “On the benefits for model regularization of a Variational formulation of GTM”, *in Proceedings of the International Joint Conference on Neural Networks (IJCNN 2008)*, in press.

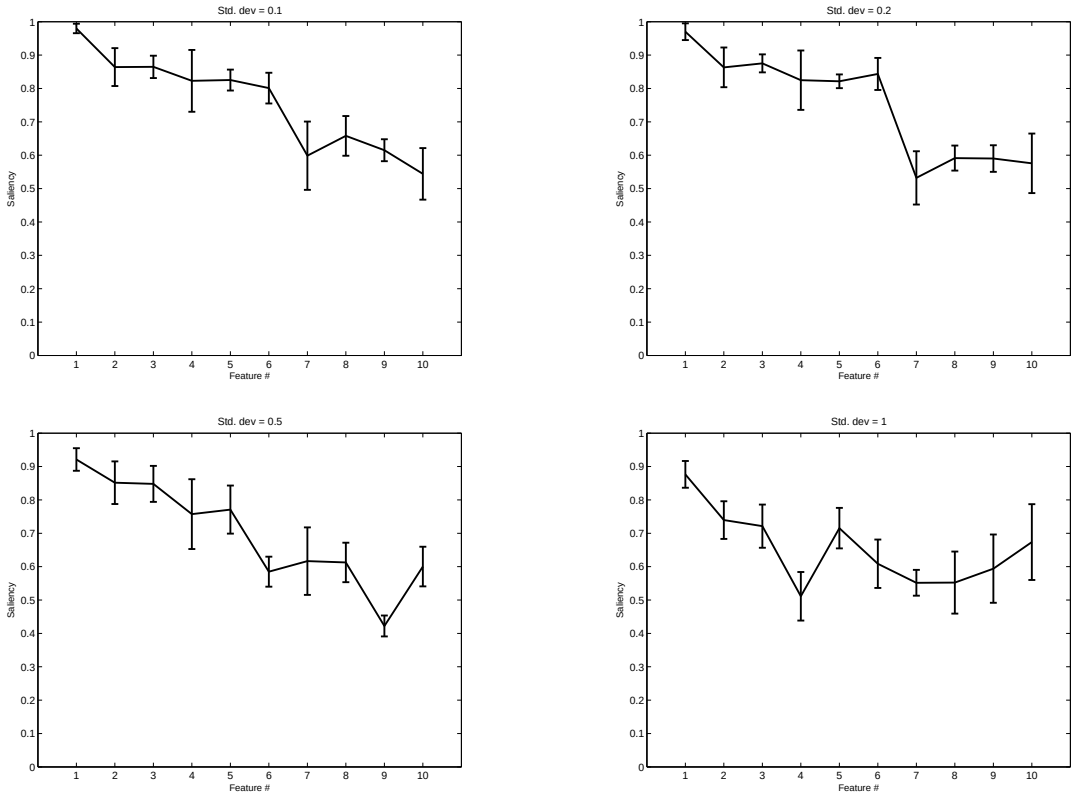

Fig. 3. Experiments with a sample of 1,000 points from *synth1*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in Fig. 1.

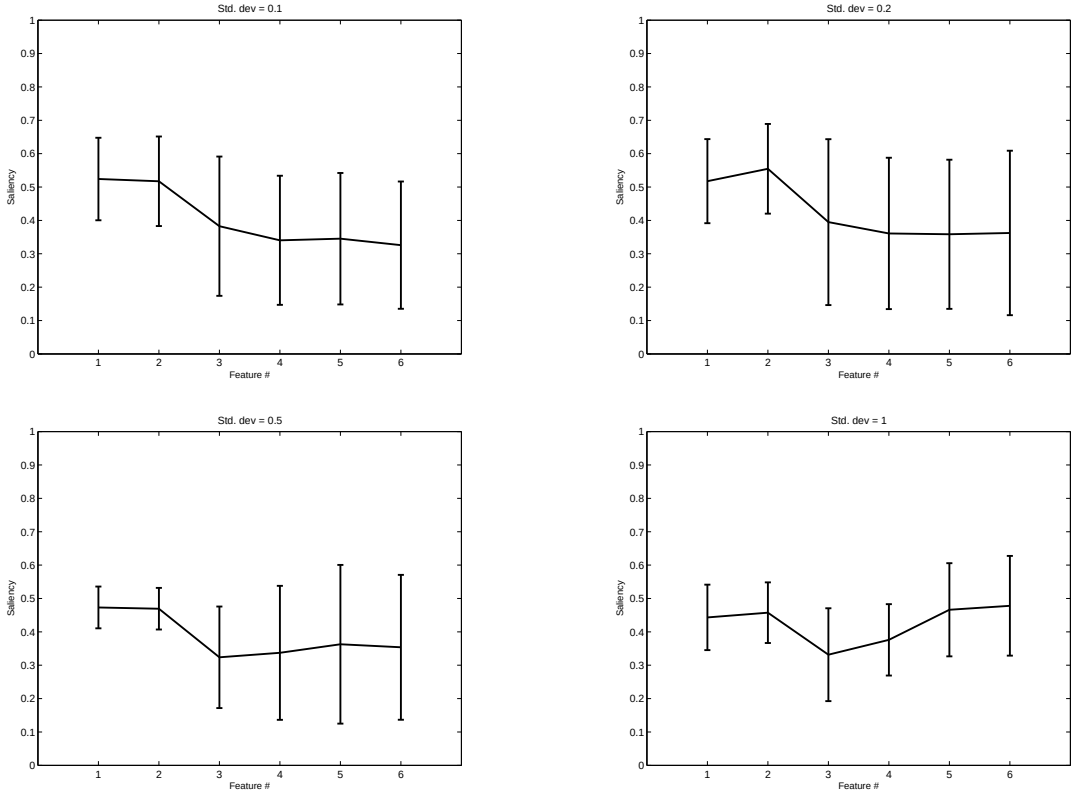

Fig. 4. Experimental results for a sample size of 1000 points from *synth2*, to which different levels of Gaussian noise (indicated in the plot titles) are added. Representation as in previous figures.

ASSESSMENT OF THE EFFECT OF NOISE ON AN UNSUPERVISED FEATURE SELECTION METHOD FOR GENERATIVE TOPOGRAPHIC MAPPING

Alfredo Vellido, Jorge S. Velazco
Department of Computing Languages and Systems (LSI), Technical University of Catalonia (UPC), Barcelona, Spain
avellido@lsi.upc.edu, e00728496@est.lsi.upc.edu

Keywords: Unsupervised Feature Selection; Feature Relevance Determination; Generative Topographic Mapping; clustering; uninformative noise.

Abstract: Unsupervised feature relevance determination and feature selection for dimensionality reduction are important issues in many clustering problems. An unsupervised feature selection method for general Finite Mixture Models was recently proposed and subsequently extended to Generative Topographic Mapping (GTM), a nonlinear manifold learning constrained mixture model for data clustering and visualization. Some of the results of a previous preliminary assessment of this method for GTM suggested that its performance may be affected by the presence of uninformative noise in the dataset. In this brief study, we test in some detail such limitation of the method.

1 INTRODUCTION

Statistical Machine Learning (SML) provides a unified principled framework for machine learning methods and helps to overcome some of their limitations. Embedding probability theory into machine learning techniques has important modeling implications. For instance, it requires modeling assumptions, including the specification of prior distributions, to be made explicit; it also automatically satisfies the likelihood principle and provides a natural framework to handle uncertainty.

An example of SML can be found in Finite Mixture Models (FMM), which are flexible and robust methods for multivariate data clustering (McLachlan and Peel, 1998). The addition of visualization capabilities would benefit these models in many application scenarios, helping to provide intuitive cues about data structural patterns. One way to endow FMM with data visualization is by constraining the mixture components to be centered in a low-dimensional manifold embedded into the multivariate data space, as in Generative Topographic Mapping (GTM) (Bishop et al., 1999). This is a non-linear, neural network-inspired manifold learning model for simultaneous data clustering and visualization.

The interpretability of the clustering results provided by GTM becomes difficult when the analyzed data sets consist of a large number of features. This limitation can be overcome with methods to estimate the ranking of the data features according to their relative relevance, leading to feature selection (FS). The research on unsupervised FS is scarce in comparison to that for supervised models, despite the fact that FS becomes a paramount issue in many clustering problems. A description of the problem in terms of a reduced subset of relevant features would improve the interpretability of the clusters obtained by unsupervised methods.

An important advance on unsupervised FS for Finite Mixture Models was presented in (Law et al., 2004) and recently extended to GTM (the FRD-GTM model) in (Vellido et al., 2006) and to one of its variants for time series analysis (FRD-GTM-TT) in (Olier and Vellido, 2006). This method was preliminarily assessed in (Vellido, 2006), where some of the results suggested that the performance of the method may be degraded by the presence of uninformative noise, which would obscure the underlying cluster structure of the data and, therefore, mislead an unsupervised feature relevance estimation method. In this brief study, we provide evidence of the limitations of

the method through controlled experiments using synthetic data.

The remaining of the paper is organized as follows. First, brief introductions to the standard Gaussian GTM and its extension for Feature Relevance Determination (FRD) are provided in section 2. This is followed, in section 3, by a description of the experimental settings and, in section 4, by a presentation and discussion of the results. The paper closes with a brief summary of conclusions.

2 FEATURE RELEVANCE DETERMINATION FOR GTM

2.1 The Standard GTM Model

The neural network-inspired GTM is a manifold learning model with sound foundations in probability theory. It performs simultaneous clustering and visualization of the observed data through a nonlinear and topology-preserving mapping from a visualization latent space in $\Re^L$ (with L being usually 1 or 2 for visualization purposes) onto a manifold embedded in the $\Re^D$ space, where the observed data reside. For each feature d, the functional form of this mapping is the generalized linear regression model $y_d(\mathbf{u},\mathbf{W}) = \sum_m^M \phi_m(\mathbf{u})\, w_{md}$, where ϕ_m is one of M basis functions, defined here as spherically symmetric Gaussians, generating the non-linear mapping from a latent vector $\mathbf{u}$ to the manifold in $\Re^D$. The matrix $\mathbf{W}$ of adaptive weights w_{md} explicitely defines this mapping.

The prior distribution of $\mathbf{u}$ in latent space is constrained to form a uniform discrete grid of K centres. A density model in data space is therefore generated for each component k of the mixture, which, assuming that the observed data set $\mathbf{X}$ is constituted by N independent, identically distributed (i.i.d.) data points $\mathbf{x}_n$, leads to the definition of a complete log-likelihood in the form:

$$L(\mathbf{W},\beta|\mathbf{X}) = \sum_{n=1}^{N} \ln\left\{ \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\beta}{2\pi}\right)^{D/2} \exp\left\{-\beta/2\|\mathbf{y}_k - \mathbf{x}_n\|^2\right\}\right\} \quad (1)$$

where $\mathbf{y}_k$ is a reference or prototype vector consisting of elements ($y_{dk} = \sum_m^M \phi_m(\mathbf{u}_k)\, w_{md}$), which are an instantiation of the generalized linear regression model described above. From Eq. (1), the adaptive parameters of the model, which are $\mathbf{W}$ and the common inverse variance of the Gaussian components, β, can be optimized by maximum likelihood (ML) using the Expectation-Maximization (EM) algorithm. Details can be found in (Bishop et al., 1999).

2.2 The FRD-GTM

In this paper, unsupervised feature relevance is understood as the likelihood of a feature being responsible for generating the data cluster structure. Therefore, relevant features will be those which better separate the natural clusters in which the data are structured. Moreover, we are interested in unsupervised feature selection methods that are suitable for clustering models that also provide data visualization. With that in mind, the FRD technique was defined for the GTM model in (Vellido et al., 2006). For the unsupervised GTM clustering model, relevance is defined through the concept of saliency.

The FRD problem was investigated for GTM in (Vellido et al., 2006). Feature relevance in this unsupervised setting is understood as the likelihood of a feature being responsible for generating the data cluster structure and it is quantified through the concept of saliency. Formally, the saliency of feature d can be defined as $\rho_d = P(\eta_d = 1)$, where η=$(\eta_1,\ldots,\eta_D)$ is a set of binary indicators that can be integrated in the EM algorithm as missing variables. A value of $\eta_d = 1$ ($\rho_d = 1$) indicates that feature d has the maximum possible relevance. According to this definition, the FRD-GTM mixture density can be written as:

$$p(\mathbf{x}|\mathbf{W},\beta,\mathbf{w}_0,\beta_0,\rho) = \sum_{k=1}^{K} \frac{1}{K} \prod_{d=1}^{D} \left\{\rho_d\, p(x_d|\mathbf{u}_k;\mathbf{w}_d,\beta) + (1-\rho_d)\, q\left(x_d|\mathbf{u}_0;w_{0,d},\beta_{0,d}\right)\right\} \quad (2)$$

where $\mathbf{w}_d$ is the vector of $\mathbf{W}$ corresponding to feature d and $\rho \equiv \{\rho_1,\ldots,\rho_D\}$. A feature d will be considered irrelevant, with *irrelevance* $(1-\rho_d)$, if $p(x_d|\mathbf{u}_k;\mathbf{w}_d,\beta) = q(x_d|\mathbf{u}_0;w_{0,d},\beta_{0,d})$ for all the mixture components k, where q is a common density followed by feature d. Notice that this is like saying that the distribution for feature d does not follow the cluster structure defined by the model. This common component requires the definition of two extra adaptive parameters: $\mathbf{w}_0 \equiv \{w_0, 1,\ldots,w_{0,D}\}$ and $\beta_0 \equiv \{\beta_{0,1},\ldots,\beta_{0,D}\}$ (so that $\mathbf{y}_0 = \phi_0(\mathbf{u}_0)\,\mathbf{w}_0$). For fully relevant ($\rho_d \to 1$) features, the common component variance vanishes:$(\beta_{0,d})^{-1} \to 0$. The parameters of the model can, once again, be optimized by ML using the EM algorithm. Detailed calculations can be found in (Vellido, 2005).

3 EXPERIMENTAL SETTINGS

The results of statistically principled models for probability density estimation, such as GTM and its variants, are bound to be affected, in one way or another, by the presence of uninformative noise in the data. Here, we assess such effects on the FRD-GTM model

described in the previous section. For that, data with very specific characteristics are required. We use synthetic sets similar to those in (Law et al., 2004) for comparative purposes.

The first synthetic set (hereafter referred to as *synth1*) is a variation on the *Trunk* data set used in (Law et al., 2004)), and was designed for its 10 features to be in decreasing order of relevance. It consists of data sampled from two Gaussians $N(\mu_1, \mathbf{I})$ and $N(\mu_2, \mathbf{I})$, where $\left(\mu_1 = 1, \frac{1}{\sqrt{3}}, \ldots, \frac{1}{\sqrt{2d-1}}, \ldots, \frac{1}{\sqrt{19}}\right)$ and $\mu_1 = -\mu_2$. We hypothesize (*H1*) that the feature relevance ranking estimated by FRD-GTM for these data will deteriorate gradually as noise is added to the 10 original features and in proportion to its level. In order to test *H1*, four increasing levels of Gaussian noise, of standard deviations 0.1, 0.2, 0.5, and 1, were added to the 10 original features of *synth1*, for a given sample size. It is also hypothesized (*H2*) that the feature relevance ranking will deteriorate as we add new noisy features and in proportion to their level of noise. In order to test *H2*, 5 and 10 dummy features consisting of Gaussian noise of standard deviations 0.1, 0.2, 0.5, and 1, were, in turn, added to the 10 original features.

The second dataset (hereafter referred to as *synth2*) consists of two features defining four neatly separated Gaussian clusters with centres located at $(0,3), (1,9), (6,4)$ and $(7,10)$; they are meant to be relatively relevant in contrast to any added noise. In a first experiment, noise of different levels was added to the first two features, while 4 extra noise features were added to those two. Several other experiments, similar to the ones devised for *synth1* were designed to further test *H2*.

The FRD-GTM parameters $\mathbf{W}$ and $\mathbf{w}_0$ were initialized with small random values sampled from a normal distribution. Saliencies were initialized at $\rho_d = 0.5, \forall d, d = 1, \ldots, D$. The grid of GTM latent centres was fixed to a square layout of 3×3 nodes (i.e., 9 constrained mixture components). The corresponding grid of basis functions ϕ_m was fixed to a 2×2 layout.

4 EXPERIMENTAL RESULTS AND DISCUSSION

The experiments outlined in the previous section aim to assess the effect of the presence of uninformative noise on the performance of FRD-GTM in the process of unsupervised feature relevance estimation.

In the experiments reported in Figure 1, four levels of Gaussian noise of increasing level were added to a sample of 1,000 points of *synth1*. The FRD-GTM is shown to behave robustly even in the presence of a substantial amount of noise, although its performance deteriorates significantly for noise of standard deviation = 1, as reflected in the breach of the expected monotonic decrease of the mean feature saliencies. It is also true that, comparing these results with those in Figure 2 (in which no noise was added to *synth1*), the most relevant feature is not so close to a saliency of 1. *H1* is, therefore, partially supported by these results.

The FRD ranking results for the second experiment, using the 10 original features of *synth1* plus 5 Gaussian noise features, are shown in Figure 2. For all levels of noise, the relevance (in the form of estimated saliency) of the original features $(1 \rightarrow 10)$ is reasonably well estimated: the saliency for the first feature is close to 1 with almost full certainty (very small vertical bars) and, overall, the expected monotonic decrease of the mean feature saliencies is preserved, although breaches of such monotonicity can also be observed. The saliencies estimated for the 5 added Gaussian noise features are regularly estimated to be small. Interestingly, the increase in the level of noise does not seem to affect the performance of the FRD method in any significant way: the differences between the saliencies of the 10 original variables and the 5 noisy ones stay roughly the same and the decreasing relevance for the 10 original variables does not vary substantially. According to these results, *H2* is not supported at this stage.

The FRD ranking results for the third experiment, using the 10 original features of *synth1* plus 10 Gaussian noise features are shown in Figure 3. Once again, and for all levels of noise, the relevance of the 10 original features shows, overall, the expected monotonic decrease of the mean feature saliencies, with some breaches of monotonicity. This time, the saliencies estimated for the 10 added Gaussian noise features are not that clearly small in comparison to those estimated for the 10 original ones. In summary, the decreasing relevance for the 10 original variables does not vary substantially, and the differences between the saliencies of the 10 original features and the 5 noisy ones stay roughly the same regardless the noise level. Nevertheless, the FRD method seems to be affected by the increase in number of the noisy features. According to these results, *H2* is only partially supported.

The FRD-GTM is shown to behave with reasonable robustness when noise is added to the first two features of *synth2*, as shown in Figure 4. As in the case of *synth1*, its performance deteriorates significantly for high levels of noise. Comparing these results with those in Figures 5 and 6 (in which no noise

was added to the first two features), the overall deterioration becomes evident. *H1* is again partially supported by these results.

The FRD ranking results for the experiments using the 2 original features of *synth2* plus either 7 or 10 Gaussian noise features are shown, in turn, in Figures 5 and 6. This is clearly a far easier problem for the FRD method. Regardless the level of noise and the number of added noisy features, FRD-GTM consistently estimates the first 2 features to be the most relevant. Furthermore, the differences between the saliencies estimated for the first 2 features and the added (7 or 10) noisy ones stay roughly the same. In contrast with the results obtained in the experiments with *synth1*, the estimated saliencies for all noisy features are low and quite similar. Our research hypothesis *H2* is not supported by these results.

5 CONCLUSION

In this paper, the effects of the presence of noise on a method of unsupervised feature relevance determination for the manifold learning GTM model, have been investigated in some detail.

The FRD-GTM has been shown to behave with reasonable robustness even in the presence of a fair amount of noise. It was first hypothesized that the feature relevance ranking would deteriorate as we add noise to the existing features and in proportion to the level of that noise. This hypothesis has found only limited experimental support. It was also hypothesized that the feature relevance ranking would deteriorate as we add extra noisy features to the existing ones and in proportion to their number and the level of noise. This second hypothesis has found little experimental support: There is only some evidence that the performance of the FRD method deteriorates as we increase the number of purely noisy features and only if the dataset is complex enough.

This relative weakness of the method in the presence of noise makes it convenient to consider possible strategies for model regularization and, therefore, future research will be devoted the design of methods for automatic and proactive model regularization to prevent or at least limit the negative effect of data overfitting on the FRD method for GTM. Some of such methods have already been designed for the standard GTM formulation (Bishop et al., 1998; Vellido et al., 2003) and could be extended to FRD-GTM. Alternatively, regularization could be accomplished through a reformulation of the GTM within a variational Bayesian theoretical framework (Olier and Vellido, 2008). Again, this could be extended to accomodate FRD.

Future research should extend the experimental design to include a wider variety of artificial data sets of different characteristics. It should also address the design of strategies for adaptive model regularization for FRD-GTM. Such kind of strategy would automatically regulate the level of map smoothing necessary to avoid the model fitting the noise in the data, i.e. data overfitting.

ACKNOWLEDGEMENTS

Alfredo Vellido is a researcher within the Ramón y Cajal program of the Spanish Ministry of Education and Science (MEC) and acknowledges funding from the MEC I+D project TIN2006-08114.

REFERENCES

Bishop, C., Svensén, M., and Williams, C. (1998). Developments of the generative topographic mapping. In *Neurocomputing*. 21(1-3), pp. 203-224.

Bishop, C., Svensén, M., and Williams, C. (1999). Gtm: The generative topographic mapping. In *Neural Computation*. 10(1), pp. 215–234.

Law, M., Figueiredo, M., and Jain, A. (2004). Simultaneous feature selection and clustering using mixture models. In *IEEE T. Pattern Anal*. 26(9), pp. 1154–1166.

McLachlan, G. and Peel, D. (1998). *Finite mixture models*. John Wiley-Sons, New York.

Olier, I. and Vellido, A. (2006). Time series relevance determination through a topology-constrained hidden markov model. In *Proc. of the 7th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2006)*. LNCS 4224, 40-47. Burgos, Spain.

Olier, I. and Vellido, A. (2008). On the benefits for model regularization of a variational formulation of gtm. In *in Proceedings of the International Joint Conference on Neural Networks (IJCNN 2008)*. in press.

Vellido, A. (2005). Preliminary theoretical results on a feature relevance determination method for generative topographic mapping. In *Technical Report LSI-05-13-R*. Universitat Politecnica de Catalunya, Barcelona, Spain.

Vellido, A. (2006). Assessment of an unsupervised feature selection method for generative topographic mapping. In *16th International Conference on Artificial Neural Networks*. LNCS 4132, 361-370. Athens, Greece.

Vellido, A., El-Deredy, W., and Lisboa, P. (2003). Selective smoothing of the generative topographic mapping. In *IEEE T. Neural Network*. 14(4), pp. 847-852.

Vellido, A., Lisboa, P., and Vicente, D. (2006). Robust analysis of mrs brain tumour data using t-gtm. In *Neurocomputing*. 69(7–9), pp. 754–768, 2006.

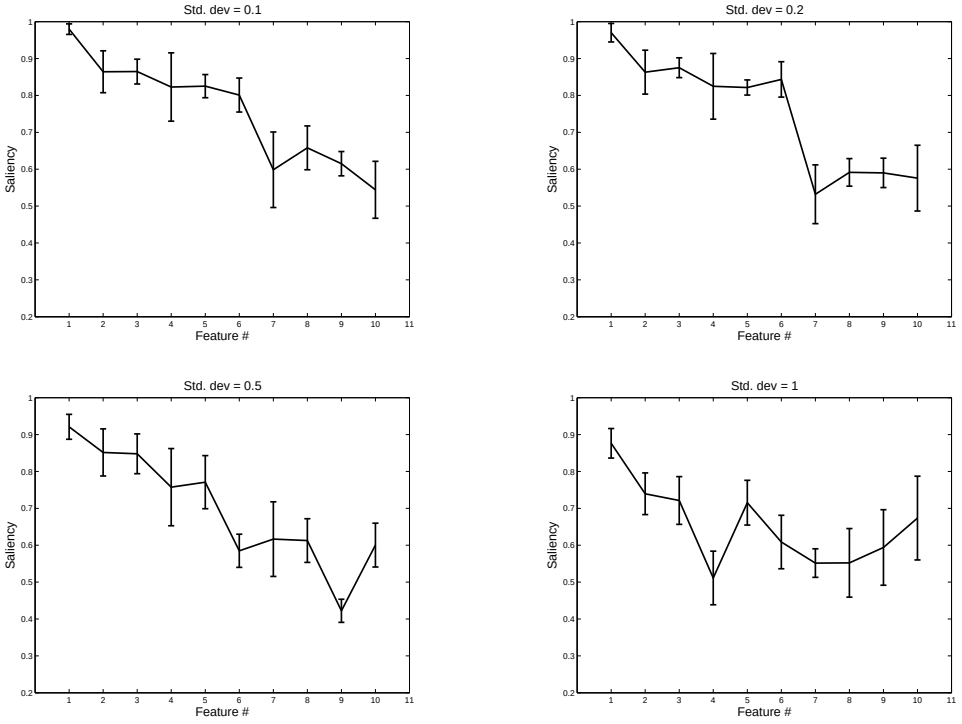

Figure 1: Experiments with a sample of 1,000 points from *synth1*, to which different levels of Gaussian noise (indicated in the plot titles) were added to the existing features. Mean saliencies ρ_d for the 10 features. The bars span from the mean minus to the mean plus one standard deviation of the saliencies over 20 runs of the algorithm.

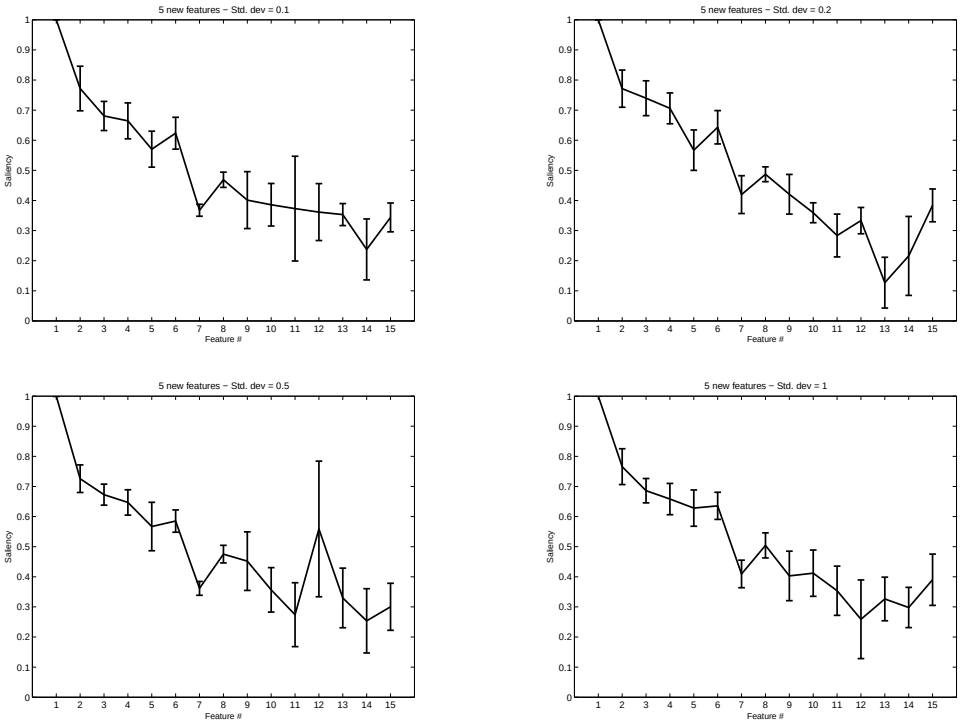

Figure 2: Experiments with a sample of 1,000 points from *synth1*, to which 5 extra noise features ($11 \rightarrow 15$) of different noise levels (indicated in the plot titles) were added. Representation as in Figure 1.

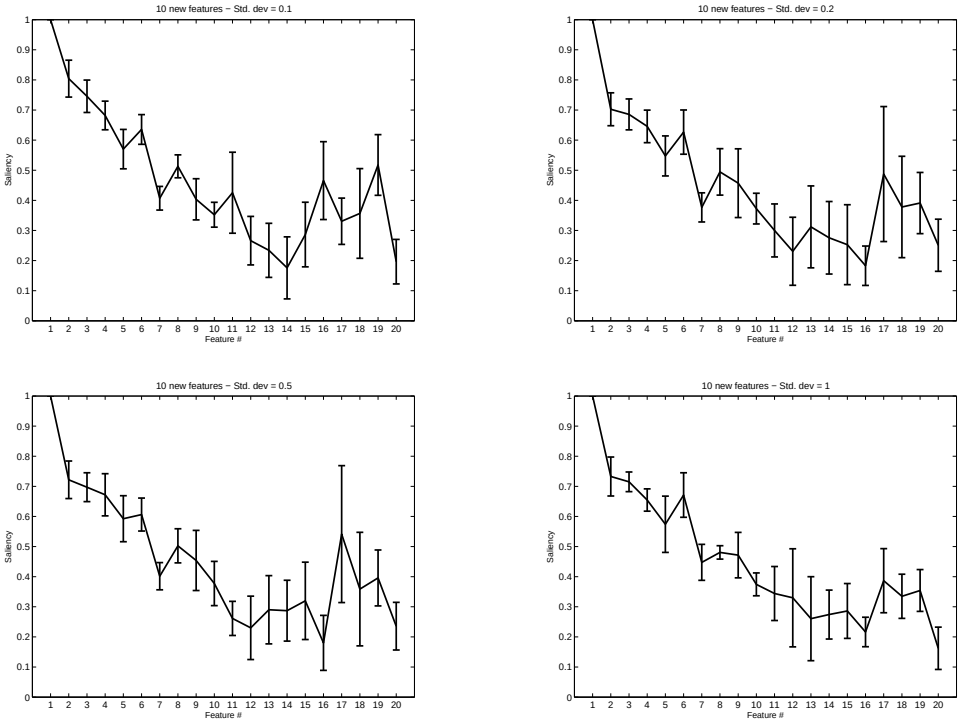

Figure 3: Experiments with a sample of 1,000 points from *synth1*, to which 10 extra noise features (11 → 20) of different noise levels (indicated in the plot titles) were added. Representation as in Figure 1.

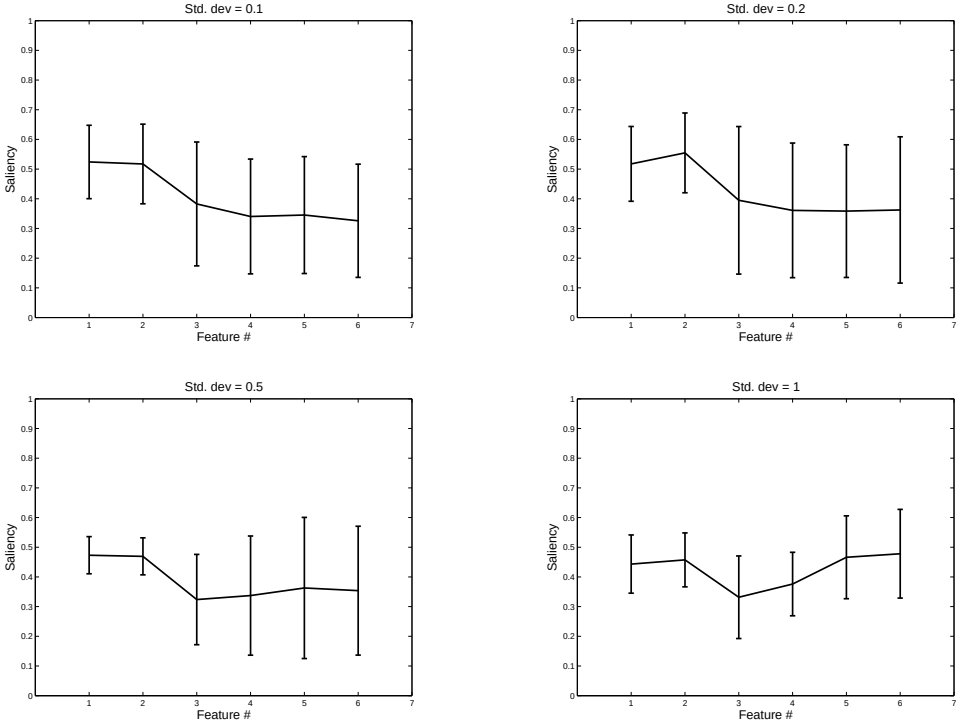

Figure 4: Experiments with a sample of 1,000 points from *synth2*, to which noise of different levels (indicated in the plot titles) were added. Four extra noise features (3 → 6) of the same noise levels were added. Representation as in previous figures.

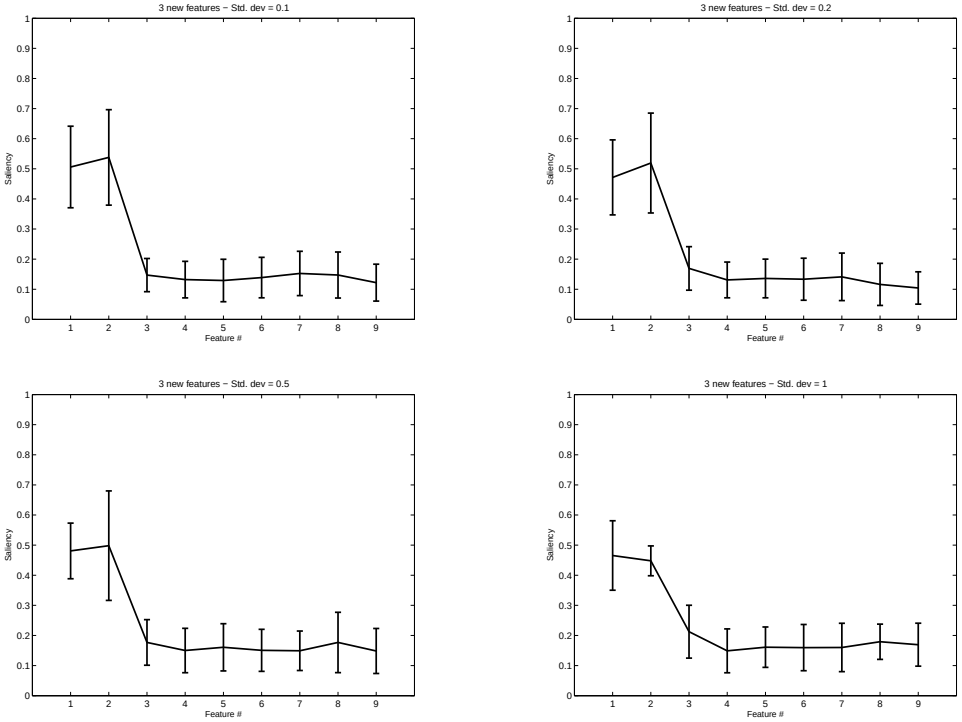

Figure 5: Experiments with a sample of 1,000 points from *synth2*, to which 7 extra noise features (3 → 9) of different noise levels (indicated in the plot titles) were added. Representation as in previous figures.

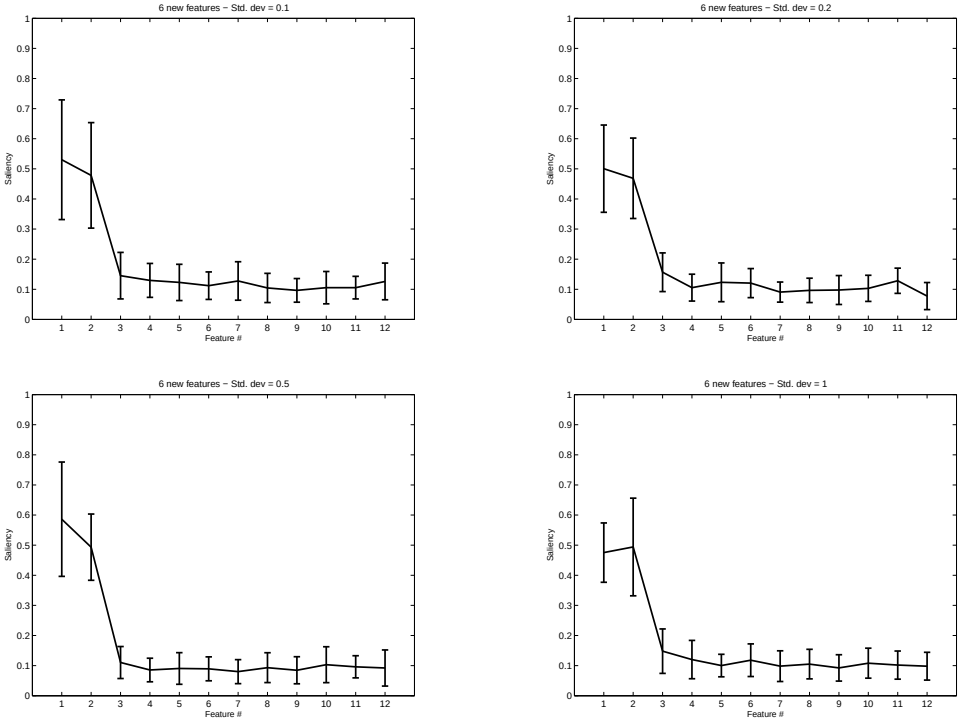

Figure 6: Experiments with a sample of 1,000 points from *synth2*, to which 10 extra noise features (3 → 12) of different noise levels (indicated in the plot titles) were added. Representation as in previous figures.

Bibliography

[1] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics, New York, 2000.

[2] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: the generative topographic mapping. *Neural Computation*, 10(1):215, 1998.

[3] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain. Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166, 2004.

[4] A. Vellido. Missing data imputation through GTM as a mixture of t-distributions. *Neural Networks*, 19(10):1624–1635, 2006.

[5] I. Olier and A. Vellido. Time series relevance determination through a topology-constrained Hidden Markov Model. *IDEAL 2006. Lecture Notes in Computer Science*, 4224:40–47, 2006.

[6] A. Vellido. Assessment of an unsupervised feature selection method for generative topographic mapping. *In Proceedings of the 16th International Conference on Artificial Neural Networks (ICANN), LNCS*, 4132:361–370, 2006.

[7] A. Vellido. Preliminary theoretical results on a feature relevance determination method for generative topographic mapping. *Technical Report LSI-05-13-R, Universitat Politècnica de Catalunya*, 2005.

[8] A. Vellido and J. Velazco. Assessment of the effect of noise on an unsupervised feature selection method for generative topographic mapping. *In 10th International Conference on Enterprise Information Systems (ICEIS 2008)*, 2008. Accepted for publication.

[9] C. M. Bishop, M. Svensén, and C. K. I. Williams. Developments of the generative topographic mapping. *Neurocomputing*, 21(1–3):203–224, 1998.

[10] A. Vellido, W. El-Deredy, and P. J. G. Lisboa. Selective smoothing of the generative topographic mapping. *IEEE Transactions on Neural Networks*, 14(4):847–852, 2003.

[11] I. Olier and A. Vellido. On the benefits for model regularization of a variational formulation of GTM. *In Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2008.

[12] I. Olier and A. Vellido. Variational GTM. *In 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07)*, 4881:77, 2007.

[13] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[14] G. J. McLachlan and D. Peel. On computational aspects of clustering via mixtures of normal and t-components. *Proceedings of the American Statistical Association (Bayesian Statistical Section)*, 2000.

[15] D. J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448, 1992.

[16] Y. A. Qi, T. P. Minka, R. W. Picard, and Z. Ghahramani. Predictive automatic relevance determination by expectation propagation. In *ICML '04: Proceedings of the 21st International Conference on Machine learning*, page 85, New York, NY, USA, 2004. ACM. ISBN 1-58113-828-5.

[17] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 2001.

[18] D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000.

[19] A. Vellido, P. Lisboa, and D. Vicente. Handling outliers and missing data in brain tumor clinical assessment using t-GTM. *Computers in Biology and Medicine*, 2006.

[20] I. Olier and J. Velazco. Basics of a variational Bayesian formulation of FRD-GTM. *Technical Report, Universitat Politècnica de Catalunya*, 2008.

[21] R. Cruz and A. Vellido. Unfolding the manifold in generative topographic mapping. *HAIS 2008 International Conference, preliminary accepted*, 2008.

[22] R. Cruz and A. Vellido. Geodesic generative topographic mapping. *IBERAMIA 2008 Conference, submitted*, 2008.