# Balancing Flexibility and Robustness in Machine Learning: Semi-parametric methods and Sparse Linear Models

José Miguel Hernández-Lobato

Computer Science Department, Universidad Autónoma de Madrid

November 5th, 2010

# Outline

# Outline

# Flexibility and Robustness in Machine Learning Methods

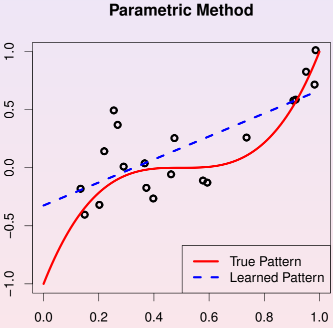
Flexibility: Capacity of a method to learn complex patterns without making strong assumptions on the actual form of such patterns.

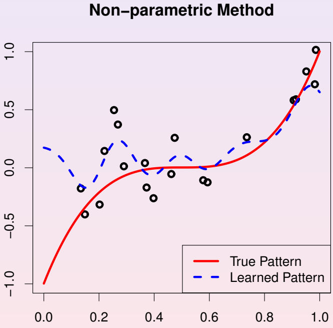
Robustness: Capacity of a method to not being affected by spurious regularities in the data, which are observed only by chance.

Flexibility and robustness are desirable, but often conflicting objectives.

# Parametric and Non-parametric Methods

Two paradigms of machine learning.
Different configurations of flexibibility and robustness.




high robustness

low flexibility

low robustness

high flexibility

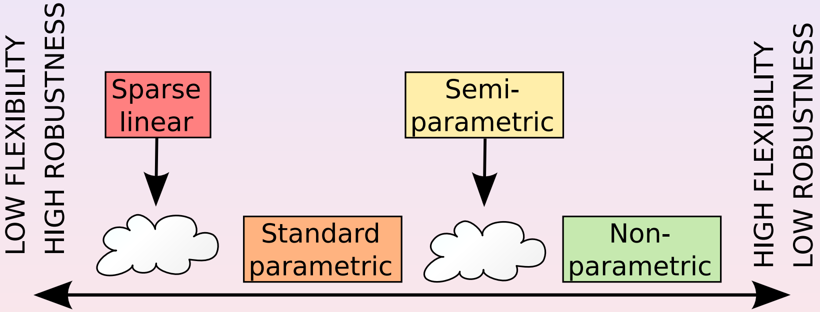
# Balancing Flexibility and Robustness

The optimal method for addressing a specific learning problem must attain the appropriate balance between flexibility and robustness.

1. In some problems, this optimal balance cannot be attained by using parametric or non-parametric approaches in isolation.
2. In other problems, even the simplest parametric methods are not sufficiently robust to provide accurate descriptions for the data.

In these situations, better results can be obtained by using a semi-parametric method (1) or assuming a sparse linear model (2).

# The Spectrum of Flexibility and Robustness



# Outline

# Semi-parametric Methods...

...include both parametric and non-parametric components in the models assumed for the data.

The parametric part of the model provides a robust description of some of the patterns present in the data.

The non-parametric component endows the model with the flexibility necessary to capture complex regularities in the data.

We propose to use semi-parametric methods for modeling:

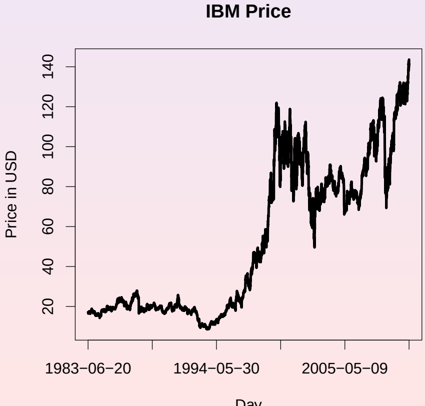
1. Time series of price changes in financial markets.
2. Non-linear dependencies between two random variables.

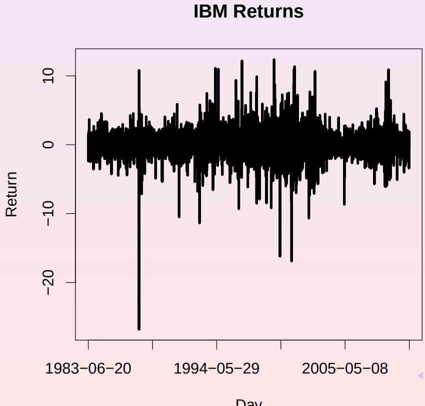
# Outline

# Time Series of Price Variations

From prices to logarithmic returns.

$$P_0, P_1, \ldots, P_n \to Y_1, \ldots, Y_n \quad \text{where} \quad Y_i = 100(\log P_i - \log P_{i-1})$$



# Semi-parametric Time Series Model for Financial Returns

$$Y_t = \mu(\mathcal{F}_{t-1};\boldsymbol{\theta}) + \sigma(\mathcal{F}_{t-1};\boldsymbol{\theta})e_t\,, \qquad t = 1, 2, \ldots, n$$

$\boldsymbol{\theta}$ is a vector of parameters.
$\boxed{e_t \sim f}$, with zero mean and unit standard deviation.
$\mathcal{F}_t$ is the information available at time $t$.

The trends $\mu(\mathcal{F}_{t-1};\boldsymbol{\theta})$ and $\sigma(\mathcal{F}_{t-1};\boldsymbol{\theta})$ are in practice simple and can be described by parametric models.

The density of the innovations $\boxed{f}$ is approximated in a non-parametric manner. This function is often complex, with non-Gaussian features such as heavy tails and negative skewness.

# Log-likelihood of the Semi-parametric Model

Given $Y_1, \ldots, Y_n$ and $\boldsymbol{\theta}$, the scaled residuals $u_1(\boldsymbol{\theta}), \ldots, u_n(\boldsymbol{\theta})$ are

$$u_t(\boldsymbol{\theta}) = [Y_t - \mu(\mathcal{F}_{t-1}; \boldsymbol{\theta})] [\sigma(\mathcal{F}_{t-1}; \boldsymbol{\theta})]^{-1} \quad t = 1, \ldots, n$$

and the corresponding log-likelihood is

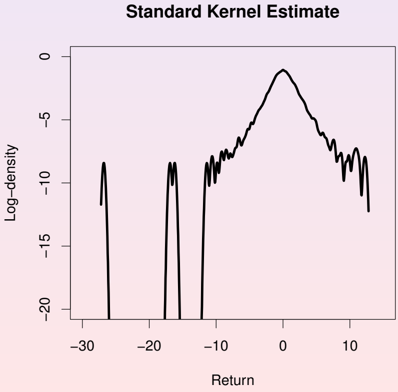
$$\mathcal{L}_n(\boldsymbol{\theta}, f \,|\, Y_1, \ldots, Y_n) = \sum_{t=1}^{n} \log f(u_t(\boldsymbol{\theta})) - \log \sigma_t(\mathcal{F}_{t-1}; \boldsymbol{\theta}), \qquad (1)$$

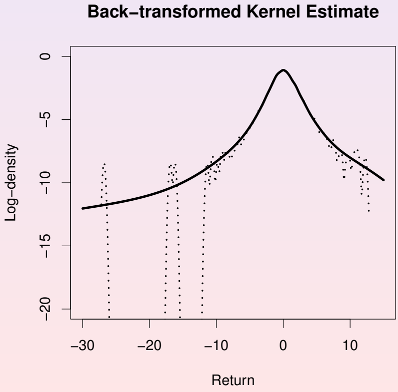
When $n \to \infty$ and $\boldsymbol{\theta}$ is hold fixed, (1) is maximized with respect to $f$ by setting $f$ to be the marginal density of $u_1(\boldsymbol{\theta}), \ldots, u_n(\boldsymbol{\theta})$.

# Back-transformed Kernel Density Estimator (BTKDE)

$$\hat{f}(x) = |g_\pi'(x)| \frac{1}{n} \sum_{i=1}^{n} K_h \left( g_\pi(X_i) - g_\pi(x) \right) \quad \text{[Wand et al. (1991)]}$$

$g_\pi(x) = \Phi^{-1}(F_\pi(x))$, $F_\pi$ is the cdf of a parametric approx.



# Iterative Algorithm for Semi-parametric Estimation

Input: a time series $Y_1, \ldots, Y_n$.
Output: a parameter vector $\hat{\boldsymbol{\theta}}$ and a density $\hat{f}$.

1. Initialize $\hat{f}$ to the standard Gaussian density.
2. $\mathcal{L}_{old} \leftarrow \infty$, $\mathcal{L}_{new} \leftarrow -\infty$.
3. while $\mathcal{L}_{new} - \mathcal{L}_{old} <$ tolerance.
    1. Update $\hat{\boldsymbol{\theta}}$ as the maximizer of $\mathcal{L}_n(\boldsymbol{\theta}, \hat{f} | Y_1, \ldots, Y_n)$.
    2. Update $\hat{f}$ as the BTKDE of the standardized $u_1(\hat{\boldsymbol{\theta}}), \ldots, u_n(\hat{\boldsymbol{\theta}})$.
    3. $\mathcal{L}_{old} \leftarrow \mathcal{L}_{new}$, $\mathcal{L}_{new} \leftarrow \mathcal{L}_n(\hat{\boldsymbol{\theta}}, \hat{f} | Y_1, \ldots, Y_n)$.
4. Return $\hat{\boldsymbol{\theta}}$ and $\hat{f}$.

[Hernández-Lobato et al. 2007]

# Experimental Evaluation on Financial Data

- ⋆ 11,665 daily returns of IBM, GM and S&P 500.
- ⋆ Trends assumed to follow an asymmetric GARCH process:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \sigma_t e_t$$
$$\sigma_t = \kappa + \alpha(|\sigma_{t-1} e_{t-1}| - \gamma \sigma_{t-1} e_{t-1}) + \beta \sigma_{t-1} \,,$$

where $\kappa > 0$, $\alpha \geq 0$, $\beta \geq 0$, $-1 < \gamma < 1$, $-1 < \phi_1 < 1$.

- ⋆ Sliding windows of size 2000. Validation on the first return out of the window. $H_0$: 9665 standard Gaussian test measurements.
- ⋆ Benchmark methods:

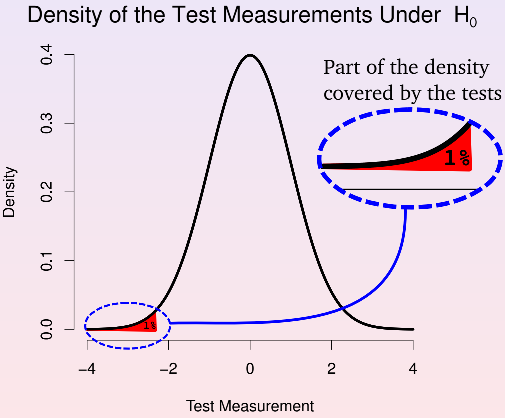
| | |
|---|---|
| MLE-NIG | [Forsberg et al. (2002)] |
| MLE-stable | [Panorska et al. (1995)] |
| SNP | [Gallant et al. (1997)] |

# Statistical Tests Described by Kerkhof et al. (2004)

Expected shortfall (ES), Value at Risk (VaR) and exceedances (Exc).

Focus on the 1% fraction of worse empirical results.

Sensitive to deviations in the loss tail: the relevant part of the distribution in risk analysis.



# Experimental Results

*p*-values of the tests described by Kerkhof et al. (2004).

| Test | Asset | SPE | MLE-NIG | MLE-stable | SNP |
|---|---|---|---|---|---|
| ES | IBM | 0.51 | 0.000001 | 0.20 | 0.0004 |
| | GM | 0.33 | 0.00004 | 0.14 | 0.001 |
| | S&P | 0.12 | 0.0005 | 0.18 | 0.000004 |
| VaR | IBM | 0.10 | 0.09 | 0.001 | 0.10 |
| | GM | 0.09 | 0.03 | 0.0002 | 0.11 |
| | S&P | 0.11 | 0.65 | 0.017 | 0.33 |
| Exc | IBM | 0.17 | 0.21 | 0.007 | 0.17 |
| | GM | 0.06 | 0.03 | 0.00008 | 0.04 |
| | S&P | 0.24 | 0.52 | 0.017 | 0.29 |

**SPE**: the proposed semi-parametric estimator. ■ *p*-value $< 0.05$.

# Conclusions Semi-parametric Time Series Models

- Back-transformed kernel density estimators (BKDE) improve the approximation of the density when the actual distribution of the data is heavy-tailed.
- An iterative algorithm (SPE) based on BKDE generates very accurate semi-parametric models of financial time series.
- SPE is a useful tool for the analysis of financial risk.

# Outline

# Copula Functions

## Sklar's Theorem

Let $(X_1, \ldots, X_d)^{\mathrm{T}} \sim F$ and let $F_1, \ldots, F_d$ be the univariate marginals of $F$. Then, there is a unique copula $C$ such that

$$F(x_1, \ldots, x_d) = C[F_1(x_1), \ldots, F_d(x_d)] \, .$$

$C$ is a distribution in $[0, 1]^d$ with uniform marginals.

$C$ captures the dependencies between $X_1, \ldots, X_d$.

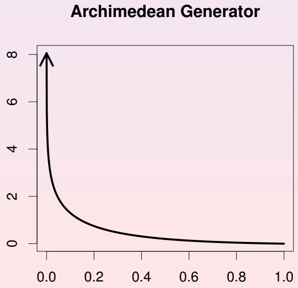
$F$ can be approximated by first, learning $F_1, \ldots, F_d$ independently and second, by learning $C$ given the estimates of the marginals.

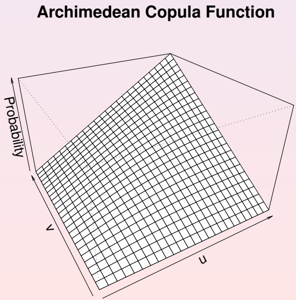
Parametric copulas may lack flexibility. Non-parametric copulas may suffer from overfitting. Solution: use semi-parametric copulas.

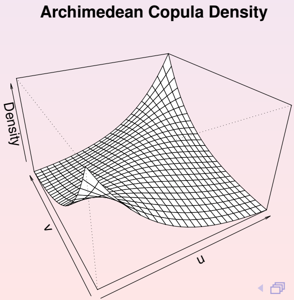
# Bivariate Archimedean Copulas

$$C(u,v) = \phi[\phi^{-1}(u) + \phi^{-1}(v)]$$

The generator $\phi^{-1} : [0,1] \to \mathbb{R}^+ \cup \{+\infty\}$ is convex, strictly decreasing, $\phi^{-1}(0) = +\infty$ and $\phi^{-1}(1) = 0$.



# Semi-parametric Bivariate Archimedean Copulas

We can obtain a semi-parametric copula model by describing $\phi^{-1}$ in a non-parametric manner. However, $\phi^{-1}$ needs to satisfy strong constraints.

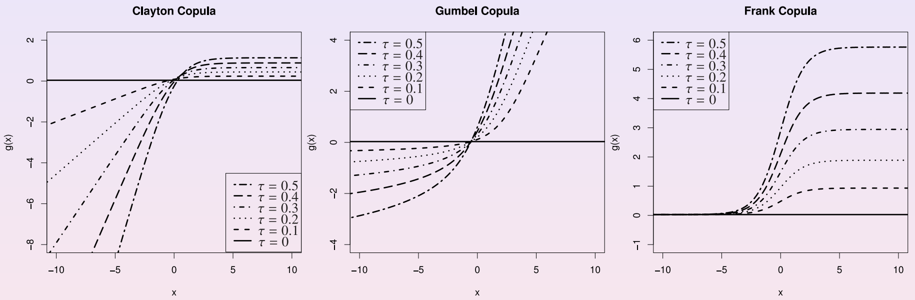
$g : \mathbb{R} \rightarrow \mathbb{R}$ is a latent function which is in a one-to-one relationship with $\phi^{-1}$ and is easier to model:

$$g(x) = \log - \frac{\phi'' \left\{ \phi^{-1}[\sigma(x)] \right\}}{\phi' \left\{ \phi^{-1}[\sigma(x)] \right\}}\,, \quad \phi^{-1}(x) = \int_x^1 \frac{1}{\int_0^y \exp\left\{ g[\sigma^{-1}(z)] \right\} dz}\, dy\,,$$

where $\sigma$ is the logistic function.

Asymptotically, $g$ behaves linearly: The asymptotic slopes of $g$ determine the level of dependence in the tails of the copula model.

# Plots of $g$ for Parametric Archimedean Copulas




These functions are well described by

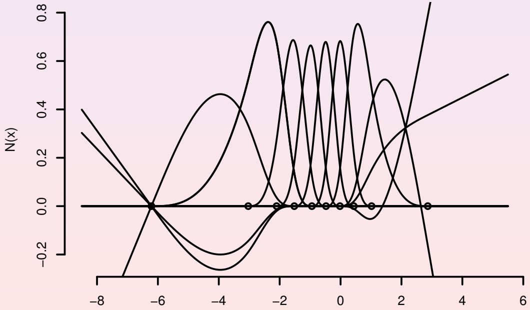
1. A central non-linear region.
2. Two asymptotically linear regions in the tails.

# Non-parametric Estimation of $g$

$g$ is described using natural cubic splines: $g_{\boldsymbol{\theta}}(x) = \sum_{i=1}^{K} \theta_i N_i(x)$

Given a sample $\mathcal{D} = \{U_i, V_i\}_{i=1}^{N}$, we maximize

$$\mathrm{PLL}(\mathcal{D}|g_{\boldsymbol{\theta}}, \beta) = \log \mathcal{L}(\mathcal{D}|g_{\boldsymbol{\theta}}) - \beta \int \left\{ g''_{\boldsymbol{\theta}}(x) \right\}^2 dx\,.$$




[Hernández-Lobato and Suárez (2009)]

# Experimental Evaluation on Financial and Rainfall Data

Conditional copula for the returns of 32 pairs of financial assets. Copula of simultaneous rainfall amounts for 32 pairs of meteorological stations.

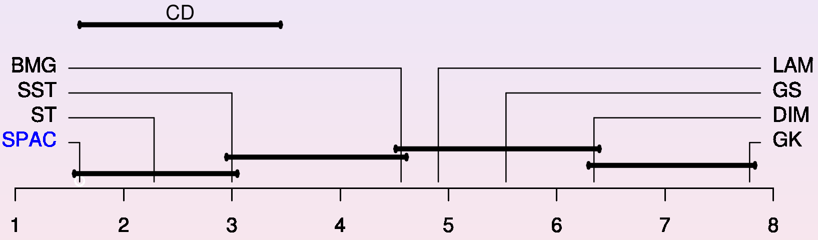
Benchmark copula estimation methods:

| | |
|---|---|
| SPAC | The proposed method. |
| LAM | Flexible Archimedean copula model [Lambert (2007)]. |
| DIM | Flexible Archimedean copula model [Dimitrova et al. (2008)]. |
| GK | Non-parametric copula based on Gaussian kernels [Fermanian et al. (2003)]. |
| BM | Copula method based on a Bayesian mixture of Gaussians. |
| ST | Parametric Student's $t$ copula. |
| GC | Parametric Gaussian copula. |
| SST | Skewed Student's $t$ copula [Demarta et al. (2005)]. |

The data are split in training and test sets with 2/3 and 1/3 of the instances. The avg. test log-likelihood is computed on each problem.

# Avg. Ranks on Financial Data and Nemenyi Test

$\alpha = 0.05$ [Demšar, J. (2006)]




*p*-values paired Wilcoxon test:

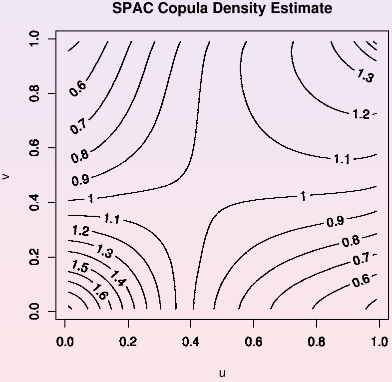
SPAC vs. ST 0.03

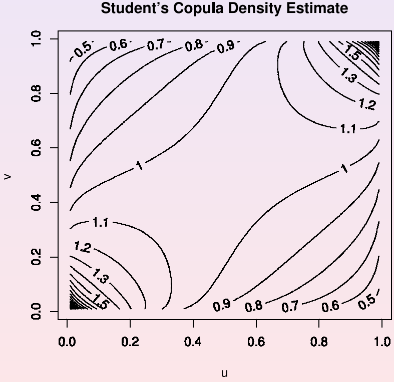
SPAC vs. SST 0.001

# Copula Density Estimates, Assets CHRW-CNP

SPAC

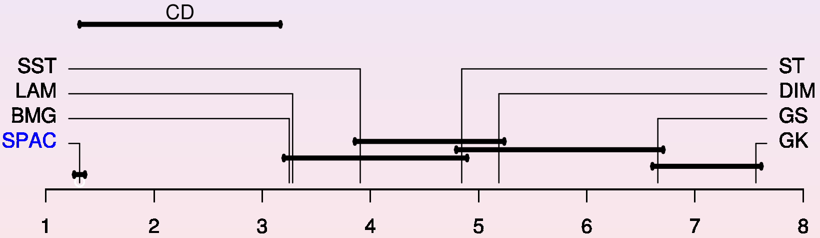
ST



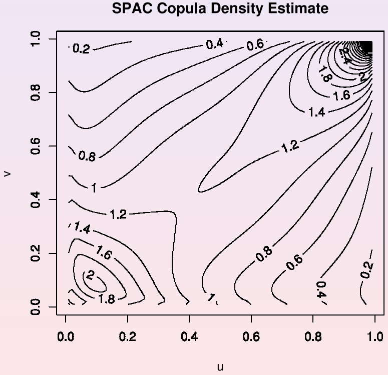
# Avg. Ranks on Precipitation Data and Nemenyi Test

$\alpha = 0.05$ [Demšar, J. (2006)]

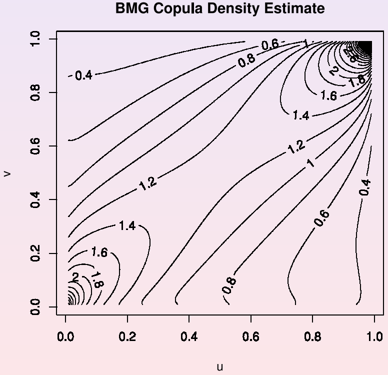


# Copula Density Estimates, Stations 30054-30253

SPAC




BMG



# Conclusions Semi-parametric Archimedean Copulas

- Expanding $g$ using a basis of natural cubic splines is a simple method (SPAC) to obtain a semi-parametric bivariate copula.
- The asymptotic slopes of $g$ determine the level of dependence in the tails of the semi-parametric dependence model.
- The good results of SPAC are explained by its capacity to model asymmetric dependencies while limiting overfitting.

# Outline

# Sparse Linear Models...

...include a few coefficients which are different from zero and many coefficients which are exactly zero.

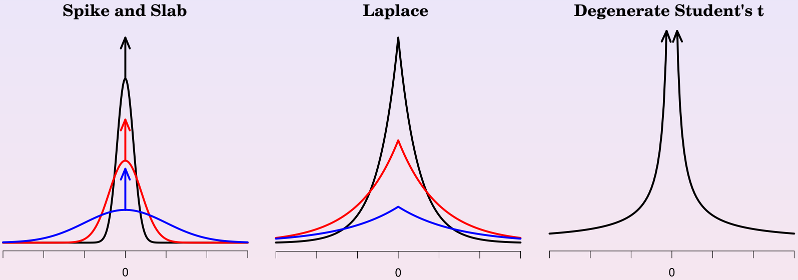
Assuming sparsity is a powerful regularization strategy that increases the robustness of the linear model at the cost of reducing its flexibility.

The resulting balance between flexibility and robustness is especially useful for addressing large $d$ and small $n$ problems.

Three main approaches for enforcing sparsity:

1. Select a small subset of features in advance.
2. Add a penalty term to the objective function.
3. Use a sparsity enforcing prior in a Bayesian approach.

# Sparsity Enforcing Priors




Selective shrinkage [Ishwaran and Rao (2005)]

We propose to use sparse linear models with spike and slab priors to address problems that belong to the large $d$ and small $n$ class.

# Outline

# The LRMSSP

The likelihood:

$$\mathcal{P}(\mathbf{y}|\mathbf{w},\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}(y_i|\mathbf{w}^\mathrm{T}\mathbf{x}_i, \sigma_0^2)\,.$$

The spike and slab prior:

$$\mathcal{P}(\mathbf{w}|\mathbf{z}) = \prod_{i=1}^{d} \left[ z_i\, \mathcal{N}(w_i|0, v_s) + (1 - z_i)\, \delta(w_i) \right], \quad \mathcal{P}(\mathbf{z}) = \prod_{i=1}^{d} \mathrm{Bern}(z_i|p_0)\,.$$

The posterior is intractable: use MCMC [George and McCulloch (1997)].

However, MCMC has often a large cost: on average $\mathcal{O}(p_0^2 d^3 k)$, $k \gg d$.

Proposed alternative: expectation propagation (EP) [Minka (2001)].

# Expectation Propagation (EP)

Approximates the posterior $\mathcal{P}(\mathbf{w}, \mathbf{z}|\mathbf{X}, \mathbf{y})$ by

$$\mathscr{Q}(\mathbf{w}, \mathbf{z}) = \prod_{i=1}^{d} \mathcal{N}(w_i|m_i, v_i)\text{Bern}(z_i|\sigma(p_i))\,,$$

where $\sigma$ is the logistic function.

Selects the parameters $m_1, \ldots, m_d$, $v_1, \ldots, v_d$, $p_1, \ldots, p_d$ by approximately minimizing $\mathrm{D}_{\mathrm{KL}}[\mathcal{P}(\mathbf{w}, \mathbf{z}|\mathbf{X}, \mathbf{y})\|\mathscr{Q}(\mathbf{w}, \mathbf{z})]$.

When $d > n$, the cost of EP is linear in $d$: $\mathcal{O}(n^2 d)$.

Expectations over $\mathscr{Q}(\mathbf{w}, \mathbf{z})$ can be computed very easily.

# Experimental Evaluation

Different regression problems with large $d$ and small $n$:

1. Reverse-engineering of transcription control networks.
2. Reconstruction of sparse signals.
3. Sentiment prediction from user-written product reviews.

Methods analyzed:

- **SS-EP** LRMSSP, EP.
- **SS-MCMC** LRMSSP, MCMC [George and McCulloch (1997)].
- **Laplace** Linear model, Laplace prior, EP [Seeger (2008)].
- **RVM** Linear model, degenerate Student's $t$ prior, type-II maximum likelihood approach [Tipping et al. (2001)].

# Experimental Results

best method.
costliest method.

## Transcription Network Reconstruction:

| | SS-MCMC | Laplace | RVM | SS-EP |
|---|---|---|---|---|
| **AUC-PR** | 19.0 | 14.9 | 14.3 | 19.4 |
| **AUC-ROC** | 75.3 | 75.1 | 64.0 | 75.7 |
| **Time** | 9041 | 4.7 | 8.7 | 7.4 |

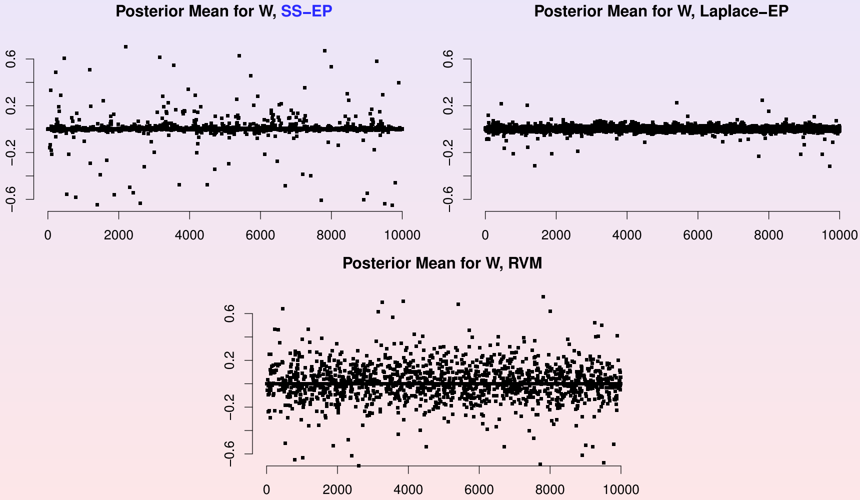
## Sparse Signal Reconstruction:

| | Non-uniform Spike Signals | | | | Uniform Spike Signals | | | |
|---|---|---|---|---|---|---|---|---|
| | SS-MCMC | Laplace | RVM | SS-EP | SS-MCMC | Laplace | RVM | SS-EP |
| **Error** | 0.19 | 0.82 | 0.19 | 0.04 | 1.03 | 0.84 | 0.66 | 0.01 |
| **Time** | 798 | 0.12 | 0.07 | 0.19 | 1783 | 0.17 | 0.12 | 0.2 |

## Sentiment Prediction:

| | Books Dataset | | | | Kitchen Appliances Dataset | | | |
|---|---|---|---|---|---|---|---|---|
| | SS-MCMC | Laplace | RVM | SS-EP | SS-MCMC | Laplace | RVM | SS-EP |
| **Error** | 1.81 | 1.84 | 2.38 | 1.81 | 1.59 | 1.64 | 1.91 | 1.59 |
| **Time** | 155,438 | 9.9 | 2.1 | 11.1 | 40,662 | 7.6 | 0.9 | 9.5 |

# Posterior Mean for **W**, Network Reconstruction Problem



# Conclusions LRMSSP

- In the LRMSSP, EP can outperform MCMC methods at a lower computational cost.
- The LRMSSP can improve the results of sparse models with Laplace and degenerate Student's $t$ priors.
- The spike and slab prior distribution has a superior selective shrinkage capacity.

# Outline

# Network of Feature Dependencies

In some classification problems with large $d$ and small $n$ there is prior information about feature dependencies.

Very often, we know that two features are likely to be either both relevant or both irrelevant for prediction.

This prior information can be encoded in an undirected network or graph $G = (V, E)$, whose nodes correspond to features and whose edges connect dependent features.

A sparse linear classifier that incorporates this prior information may improve its predictive performance.

# A Network Based Sparse Bayesian Classifier (NBSBC)

The information in $G$ can be included into a sparse linear classifier with spike and slab priors by using a Markov random field as the prior for $\mathbf{z}$:

$$\mathcal{P}(\mathbf{z}|G,\alpha,\beta) = \frac{1}{Z}\exp\left\{10z_0 + \alpha\sum_{i=1}^{d} z_i\right\} \boxed{\exp\left\{\beta \sum_{\{j,k\}\in E} z_j z_k\right\}}\,.$$

Let $\Theta$ be the Heaviside step function. Then, the classification likelihood is

$$\mathcal{P}\left(\mathbf{y}|\mathbf{w},\epsilon,\mathbf{X}\right) = \prod_{i=1}^{n}\left[\epsilon\left(1-\Theta\left(y_i\mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right)\right) + (1-\epsilon)\Theta\left(y_i\mathbf{w}^{\mathsf{T}}\mathbf{x}_i\right)\right]$$

and the prior for the noise in the class labels is $\mathcal{P}(\epsilon) = \mathrm{Beta}(\epsilon|a_0,b_0)$.

EP is used for approximate inference [Hernández-Lobato et al. (2010)]

# Experimental Evaluation of NBSBC

Different classification problems with a network of features $G$:

1. English phonemes (aa vs. ao).
2. Handwritten digits (7 vs. 9) (background noise).
3. Precipitation amounts (positive vs. zero).
4. Metastasis-free survival time (larger vs. shorter).

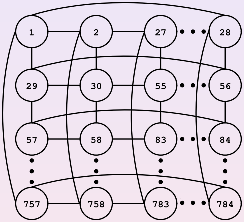
Methods analyzed:

- NBSBC The proposed method.
- SBC The proposed method with no network info ($\beta = 0$).
- NBSVM The network-based support vector machine [Zhu et al. (2009)].
- GL Logistic regression with a graph lasso penalty [Jacob et al. (2009)].
- SVM The standard support vector machine.

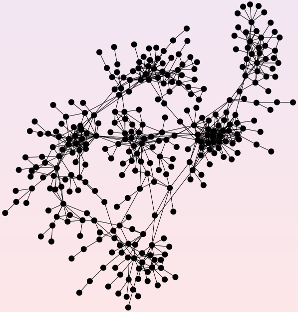
# Networks of Features for each Problem

**English Phonemes**


**Handwriten digits**




**Metastasis-free survival time**



**Precipitation amounts**

# Experimental Results

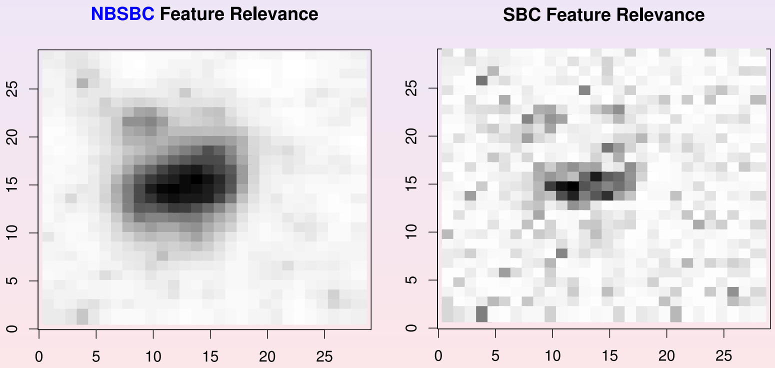
Average test error for each method:

| | SVM | NBSVM | GL | SBC | NBSBC |
|---|---|---|---|---|---|
| Phonemes | 20.66 | 20.24 | 20.55 | 20.19 | 19.48 |
| Digits | 10.32 | 10.23 | 11.18 | 9.18 | 8.35 |
| Precipitation | 38.12 | 36.69 | 32.31 | 35.16 | 33.17 |
| Metastasis | 33.20 | 34.67 | 36.31 | 32.95 | 32.23 |

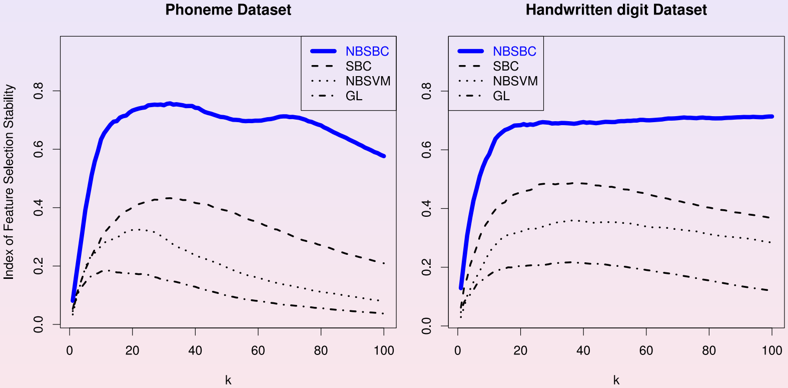
best performing method.

# Feature Relevance for NBSBC and SBC in Digits

Posterior probabilities of the latent variables $z_0, \ldots, z_d$:



# Stability of the Different Methods in Phonemes and Digits




Agreement between feature rankings in the different train/test episodes.

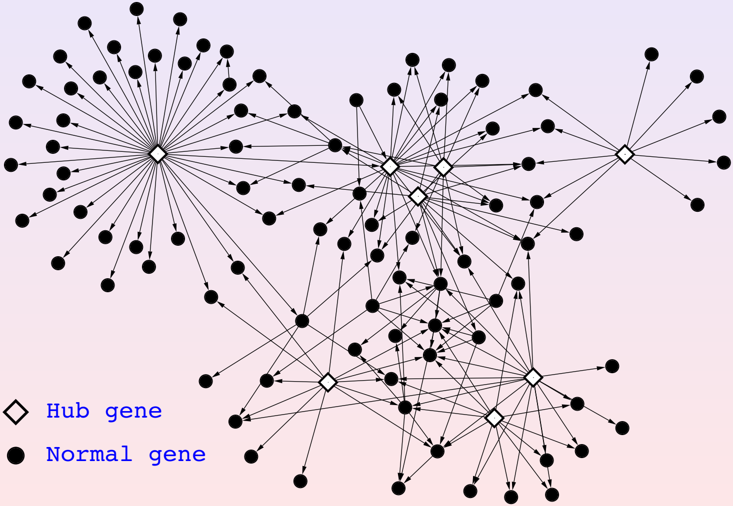
[Kuncheva (2007)]

# Conclusions NBSBC

- Taking into account dependencies between features can improve the predictive performance of a sparse linear model.
- These dependencies can be incorporated into a model with spike and slab priors by using a Markov random field.
- NBSBC is very robust and stable against small perturbations of the training set.

# Outline

# Regulators in Transcription Networks are Highly Connected



# A Hierarchical Sparse Linear Model for Gene Regulation

The expression at $t+1$ is a linear function of the expression at $t$:

$$\mathbf{x}_{t+1} = \mathbf{W}\mathbf{x}_t + \text{Gaussian noise}\,.$$

The prior for $\mathbf{W}$ is spike and slab conditioning to $\mathbf{Z}$.

A hierarchical prior for $\mathbf{Z}$ encodes domain knowledge about hubs:

$$\mathcal{P}(\mathbf{Z}|\mathbf{r}) = \prod_{i=1}^{d} \prod_{j=1,\, j\neq i}^{d} \left[r_j \text{Bern}(z_{ij}|p_1) + (1-r_j)\text{Bern}(z_{ij}|p_0)\right] \prod_{k=1}^{d} (1 - z_{kk})\,,$$

where $\mathbf{r} = (r_1, \ldots, r_d)^\text{T}$ is a binary latent vector that indicates which genes are regulators and $p_1 > p_0$.

The posterior of $r_i = 1$ gives the probability that gene $i$ is a regulator.

EP for approximate inference. [Hernández-Lobato et al. (2008)]

# Experiments on Real Microarray Data (Yeast)

cdc dataset [Spellman et al. (1998)]. 751 genes, 23 measurements.

Among the top ten genes:

| Rank | Gene | Annotation |
|---|---|---|
| 1 | YLR098c | DNA binding transcriptional activator |
| 2 | YOR315w | Putative transcription factor |
| ... | ... | ... |
| 6 | YLR095w | Transcription elongation |
| ... | ... | ... |

4% of the yeast genome is associated with transcription. Thus, the probability of finding 3 regulators among 10 genes by chance is 0.0058.

# Experiments on Real Microarray Data (Malaria Parasite)

3D7 dataset [Linás et al. (2006)]. 751 genes, 53 measurements.

Among the top ten genes:

| Rank | Gene | Annotation or BLASTP hits |
|---:|---|---|
| 1 | PFC0950c | 25% identity to GATA TF in Dictyostelium |
| 2 | PF11_0321 | 25% identity to putative WRKY TF in Dictyostelium |
| ... | ... | ... |
| 5 | PFD0175c | 32% identity to GATA TF in Dictyostelium |
| 6 | MAL7P1.34 | 35% identity to GATA TF in Dictyostelium |
| ... | ... | ... |
| 10 | MAL13P1.14 | DEAD box helicase |

# Conclusions Discovering Regulatory Genes

- Regulators are usually highly connected nodes (hubs) in transcription control networks.
- Regulators can be identified from microarray data by using a linear model with a hierarchical spike and slab prior.
- Experiments with simulated and actual microarray data validate the proposed approach.

# Outline

# Future Work

Semi-parametric methods:

1. Study the asymptotic convergence of the iterative algorithm.
2. Analyze alternative transformation functions for BKDE.
3. Extend semi-parametric copulas to higher dimensions.
4. SPAC for modeling time-series of rainfall measurements.

Sparse linear models:

1. Apply the LRMSSP to active learning problems.
2. Spike and slab priors in recommender systems.
3. Spike and slab priors for multi-task learning.
4. Extend the hierarchical model for gene regulation to incorporate information about DNA sequence.

# References I


- Wand, M. P., Marron, J. S., and Ruppert, D. (1991). Transformations in density estimation. with discussion and a rejoinder by the authors. Journal of the American Statistical Association, 86(414):343-361.
- Hernández-Lobato, J. M., Hernández-Lobato, D., and Suárez, A. (2007). Garch processes with non-parametric innovations for market risk estimation. ICANN, Volume 2, pages 718-727, Springer.
- Kerkhof, J. and Melenberg, B. (2004). Backtesting for risk-based regulatory campital. Journal of Banking and Finance, 28(8):1845-1865.
- Forsberg, L. and Bollerslev, T. (2002). Bridging the gap between the distribution of realized (ECU) volatility and ARCH modelling (of the euro): the GARCH-NIG model. Journal of Applied Econometrics, 17:535-548.
- Panorska, A. K., Mittnik, S., and Rachev, S. T. (1995). Stable GARCH models for financial time series. Applied Mathematics Letters, 8(5):33-37.
- Gallant, A. R., Hsieh, D., and Tauchen, G. (1997). Estimation of stochastic volatility models with diagnostics. Journal of Econometrics, 81(1):159-192.
- Hernández-Lobato, J. M. and Suárez, A. (2009). Modeling dependence in financial data with semiparametric archimedean copulas. In International Workshop AMLCF, London UK.
- Lambert, P. (2007). Archimedean copula estimation using Bayesian splines smoothing techniques. Computational Statistics & Data Analysis, 51:6307-6320.
- Dimitrova, D. S., Kaishev, V. K., and Penev, S. I. (2008). GeD spline estimation of multivariate Archimedean copulas. Computational Statistics & Data Analysis, 52(7):3570-3582.
- Fermanian, J. and Scaillet, O. (2003). Nonparametric estimation of copulas for time series. The Journal of Risk, 5(4):25-54.
- Demarta, S. and McNeil, A. J. (2005). The t copula and related copulas. International Statistical Review, 73(1):111-129.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7:1-30.

# References II


- Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. The Annals of Statistics, 33(2):730-773.
- George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. Statistica Sinica, 7(2):339-373.
- Minka, T. (2001). A Family of Algorithms for Approximate Bayesian Inference. PhD thesis, MIT.
- Seeger, M., Nickisch, H., and Schlkopf, B. (2010). Optimization of k-space trajectories for compressed sensing by Bayesian experimental design. Magnetic Resonance in Medicine, 63(1):116-126.
- Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. The Journal of Machine Learning Research, 1:211-244.
- Hernández-Lobato, J. M., Hernández-Lobato, D., and Suárez, A. (2010). Network-based sparse Bayesian classification. Pattern Recognition. In Press.
- Zhu, Y., Shen, X., and Pan, W. (2009). Network-based support vector machine for classification of microarray samples. BMC Bioinformatics, 10(Suppl 1):S21.
- Jacob, L., Obozinski, G., and Vert, J. (2009). Group lasso with overlap and graph lasso. In ICML 2009, pages 433-440.
- Kuncheva, L. I. (2007). A stability index for feature selection. In Proceedings of the 25th conference on Proceedings of the 25th IASTED International Multi-Conference: artificial intelligence and applications, pages 390-395. ACTA Press.
- Hernández-Lobato, J. M., Dijkstra, T., and Heskes, T. (2008). Regulator discovery from gene expression time series of malaria parasites: a hierachical approach. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, Advances in Neural Information Processing Systems 20, pages 649-656. MIT Press.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher., B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell, 9(12):3273-3297.
- Llinás, M., Bozdech, Z., Wong, E. D., Adai, A., and DeRisi, J. L. (2006). Comparative whole genome transcriptome analysis of three Plasmodium falciparum strains. Nucleic Acids Research, 34(4):1166-1173.

# Thank you for your attention!