Support Vector Machine with Missing Class Identifiers

Introduction

A motivating example: Complicated Grief (CG)

Acute grief: a normal human response to the loss of a loved one, usually dissipates with time.

Complicated grief (CG): an intense and long-lasting form of grief that takes over a person's life.

Studied since mid 90's, prevalence about 10% in bereaved people.

Those suffering from CG often exhibit symptoms of (Shear et al. 2011):

- strong yearning for the person who died
- frequent intrusive thoughts of the loved one
- feelings of intense loneliness or emptiness
- a feeling that life without the person has no purpose or meaning





Scientific goal

- To use psychiatric assessment information ($\boldsymbol{X}$) to predict CG disease (D).

Feature variables $\boldsymbol{X}$

Inventory of Complicated Grief (ICG; Prigerson et al. 1995)

1. Preoccupation with the person who died
2. Memories of the person who died are upsetting
3. The death is unacceptable
4. Longing for the person who died
5. Drawn to places and things associated with the person who died
6. Anger about the death
7. Disbelief
8. Feeling stunned or dazed
9. Difficulty trusting others
10. Difficulty caring about others
11. Avoidance of reminders of the person who died
12. Pain in the same area of the body
13. Feeling that life is empty
14. Hearing the voice of the person who died
15. Seeing the person who died
16. Feeling it is unfair to live when the other person has died
17. Bitter about the death
18. Envious of others
19. Lonely

Practical challenge

- However, CG is never well defined.
- There has even been some debate regarding whether CG should be a disease:
  - In the Diagnostic Statistical Manual of Mental Disorders DSM-IV (1994), CG is not explicitly considered a mental disorder.
  - But body of evidence leads to inclusion of CG to DSM-5, Section 3 (due out May 2013).
- Challenge: A prediction problem but in the absence of disease state!

Salvaged by disease-informative markers

- There often exist functional markers as the consequences of disease.
- One important marker for CG is

Work and Social Adjustment Scale (Mundt et al. 2002)

Rate each of the following questions on a 0 to 8 scale: 0 indicates no impairment at all and 8 indicates very severe impairment.

1. Because of my [disorder], my ability to work is impaired. 0 means not at all impaired and 8 means very severely impaired to the point I can't work.

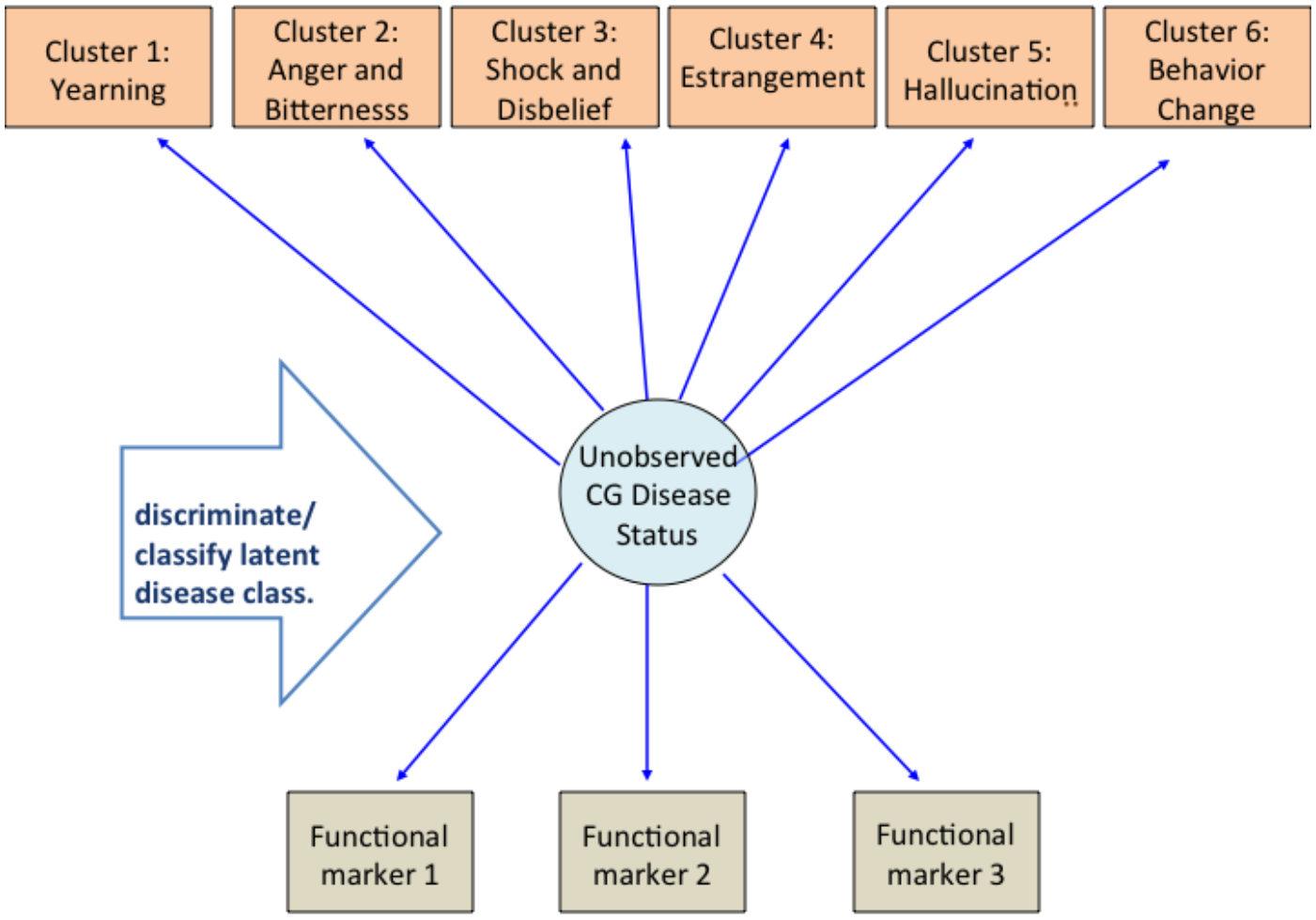
2. Because of my [disorder], my home management (cleaning, tidying, shopping, cooking, looking after home or children, paying bills) is impaired. 0 means not at all impaired and 8 means very severely impaired.

3. Because of my [disorder], my social leisure activities (with other people, such as parties, bars, clubs, outings, visits, dating, home entertainment) are impaired. 0 means not at all impaired and 8 means very severely impaired.

4. Because of my [disorder], my private leisure activities (done alone, such as reading, gardening, collecting, sewing, walking alone) are impaired. 0 means not at all impaired and 8 means very severely impaired.

5. Because of my [disorder], my ability to form and maintain close

Diagram of prediction problem



Methods

Overview of proposed approach

Goal: to construct an optimal decision rule to classify presence/absence of a disorder:

- Absence of a gold standard disease status
- Observe feature variables, $\boldsymbol{X}$
- Observe external disease-informative markers, $\boldsymbol{Z}$

Overview of proposed approach

Classification, variable/item selection in the presence of missing data (missing class labels).

- Large-margin based classifiers
- Pseudo-EM: introduce EM algorithm for missing data to margin-based classification
  - Distinction: do NOT model distribution of disease status and feature variables
  - Pseudo-likelihood naturally based on the loss function.

Borrow information from external informative markers to guide classification.

Large margin based classification with known labels

Let $D_i = 1$: diseased; $D_i = -1$: non-diseased; $g(\boldsymbol{x})$ separating boundary

Margin based loss subject to a penalty

$$\arg \min_{g \in \mathcal{H}_K} \left[\sum_i \mathcal{L}\{D_i g(\boldsymbol{X}_i)\} + \frac{\lambda_n}{2} ||g||^2 \right],$$

- SVM loss: $(1-u)_+$
- Variant of SVM loss: $(1-u)_+^q$
- $\psi-$loss: $\mathcal{L}(u) = \psi(u), \psi(u) = 1 - \text{sign}(u)$ if $u \geq 1$ or $u < 0$; and $\psi(u) = 2(1-u)$ otherwise

Assumptions

Key assumptions:

- $\boldsymbol{Z}_i | D_i = d \sim MVN(\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)$
- Disease-informative marker has distinct population mean and known direction
- For illustrative purposes, $\boldsymbol{Z}_i$ and $\boldsymbol{X}_i$ are conditionally independent given D_i (can be relaxed)
- Other technical assumptions for theoretical proof

Methods: SVM-EM

Penalized pseudo-likelihood:

$$\prod_i \exp\left[-\mathcal{L}\{D_i g(\boldsymbol{X}_i)\} - \frac{\lambda_n}{2}||g||^2\right].$$

Penalized pseudo-log-likelihood for complete data $(\boldsymbol{Z}_i, D_i, \boldsymbol{X}_i)$ (up to some constant):

$$\sum_i \log\left[f(\boldsymbol{Z}_i|D_i = d; \boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d)\right] - \sum_i \left[\mathcal{L}\{D_i g(\boldsymbol{X}_i)\} + \frac{\lambda_n}{2}||g||^2\right].$$

Pseudo-EM algorithm accounting for missing D_i (SVM-EM)

Methods: SVM-EM

E-step:

$$\begin{aligned} w_i^{(m)} &= E(D_i = 1|\text{observed}) \\ &= \frac{f(\boldsymbol{Z}_i; \widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\Sigma}}_1)\Pr^{(m-1)}(D_i = 1|\boldsymbol{X}_i)}{\sum_{d\in\{-1,1\}} f(\boldsymbol{Z}_i; \widehat{\boldsymbol{\mu}}_d, \widehat{\boldsymbol{\Sigma}}_d)\Pr^{(m-1)}(D_i = d|\boldsymbol{X}_i)}, \end{aligned}$$

Pseudo-probabilities

$$\Pr^{(m-1)}(D_i = d|\boldsymbol{X}_i) = \frac{p[d, g^{(m-1)}(\boldsymbol{X}_i)]}{p[1, g^{(m-1)}(\boldsymbol{X}_i)] + p[-1, g^{(m-1)}(\boldsymbol{X}_i)]},$$

where

$$p[d, g(\boldsymbol{x})] = \exp[-\mathcal{L}\{dg(x)\}].$$

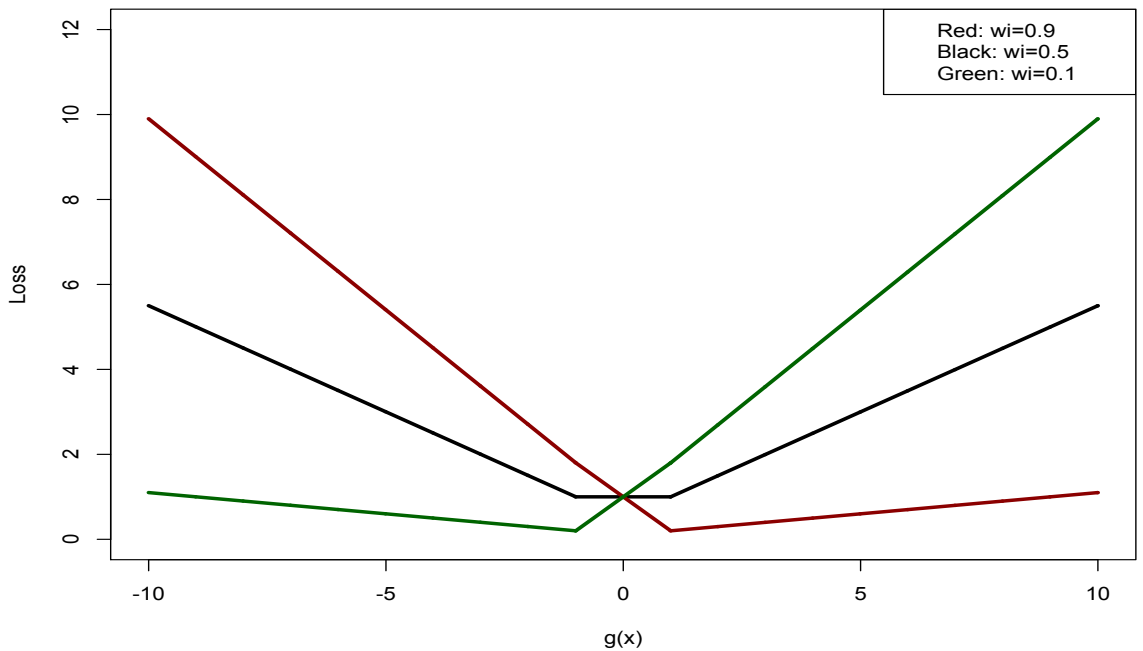
Methods: SVM-EM

For $\boldsymbol{\mu}_d, \boldsymbol{\Sigma}_d$: Gaussian mixture model

M-step for $g(\cdot)$:

$$\min_{g \in \mathcal{H}_K} \left(\sum_i \left[w_i^{(m)} \mathcal{L}\{g(X_i)\} + (1 - w_i^{(m)}) \mathcal{L}\{-g(X_i)\} \right] + \frac{\lambda_n}{2} ||g||^2 \right)$$

Figure: Illustration under the SVM loss, $\mathcal{L}(u) = (1-u)_+$



Methods: SVM-EM

Under linear decision boundary and SVM loss:

$$\arg\min_{\beta_0,\boldsymbol{\beta}} \left(\sum_i \xi_i + \frac{\lambda_n}{2} \|\boldsymbol{\beta}\|^2 \right)$$

$$u_i^{(m-1)}(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}_i) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \cdots, n,$$

$$u_i^{(m)} = E(D_i|\text{observed}) = 2w_i^{(m)} - 1.$$

Dual form:

$$\max \left(\sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j u_i^{(m)} u_j^{(m)} \boldsymbol{X}_i^T \boldsymbol{X}_j \right),$$

$$0 \leq \alpha_i \leq C_n, \quad \sum_{i=1}^n \alpha_i \text{sign}\{u_i^{(m)}\} = 0, \quad i = 1, ..., n.$$

Methods: SVM-EM

Computation extremely easy!

- Re-scaled $\widetilde{\boldsymbol{X}}_i = |u_i^{(m)}|\boldsymbol{X}_i$, $\widetilde{D}_i = \text{sign}\{u_i^{(m)}\}$ and solve regular SVM

$$\max \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \widetilde{D}_i \widetilde{D}_j \widetilde{\boldsymbol{X}}_i^T \widetilde{\boldsymbol{X}}_j \right),$$

$$0 \leq \alpha_i \leq C_n, \sum_{i=1}^{n} \alpha_i \widetilde{D}_i = 0, \quad i = 1, ..., n.$$

- Adapt to nonlinear boundary by kernel trick

Methods: SVM-EM

Variable selection while accommodating correlation among $\boldsymbol{X}$:

$$\min_{b,\boldsymbol{\beta}} \Big[\sum_i \Big\{ w_i^{(m)} \mathcal{L}(b + \boldsymbol{\beta}^T \boldsymbol{X}_i) + (1 - w_i^{(m)}) \mathcal{L}(-b - \boldsymbol{\beta}^T \boldsymbol{X}_i) \Big\}$$

$$+\lambda_{1n} ||\boldsymbol{\beta}||_1 + \frac{\lambda_{2n}}{2} ||\boldsymbol{\beta}||^2 \Big].$$

Computation: penalized SVM (Zhu et al. 2004; Becker et al. 2009)

Asymptotic Results

Theoretical property I:

Convergence of the pseudo-EM algorithm:

Theorem

Let $\widetilde{p}[d, g(\cdot)] = \exp\left[-\mathcal{L}\{dg(\cdot)\} - p_{\lambda_n}(g)\right]$, *where* $p_{\lambda_n}(g)$ *is a penalty function, and define*

$$\mathcal{Q}_n(g) \equiv \sum_{i=1}^{n} \log\left\{\sum_d \widetilde{p}[d, g(\boldsymbol{X}_i)] f(\boldsymbol{Z}_i | D_i = d; \widehat{\boldsymbol{\mu}}_d, \widehat{\boldsymbol{\Sigma}}_d)\right\}.$$

Then $\mathcal{Q}_n(g)$ *is non-decreasing after each iteration in the EM algorithm. Furthermore, the value of* $\mathcal{Q}_n$ *does not increase if and only if the decision rule based on* $g(\cdot)$ *does not change after an iteration.*

Theoretical property II:

Fisher consistency under a general boundary:

Theorem

Let $g_0(\boldsymbol{x})$ *be* $p_1(\boldsymbol{x}) - p_{-1}(\boldsymbol{x})$ *where* $p_d(\boldsymbol{x}) = Pr(D = d|\boldsymbol{X} = \boldsymbol{x})$. *Then*

$$\text{sign}[g^*(\boldsymbol{x})] = \text{sign}[g_0(\boldsymbol{x})],$$

where $g^*(\cdot)$ *is the limit of* $\widehat{g}(\cdot)$.

Theoretical properties III:

Paremeter consistency and variable selection consistency under linear boundary:

Theorem

Under a linear decision rule, $g(\boldsymbol{x}) = b_0 + \boldsymbol{\beta}^T \boldsymbol{x}$, *and technical conditions on* $p_{\lambda_n}(|\boldsymbol{\beta}|)$,
(a) $\widehat{\boldsymbol{\beta}}$ *is consistent;*
(b) with probability tending to 1, $\text{sign}(\widehat{\beta}_k) = \text{sign}(\beta_{0k})$, *where* $\widehat{\beta}_k$ *and* β_{0k} *are the* k*th component of* $\widehat{\boldsymbol{\beta}}$ *and* $\boldsymbol{\beta}_0$, *respectively;*
(c) $\sqrt{n}(\widehat{\boldsymbol{\beta}}^1 - \boldsymbol{\beta}_0^1)$ *converges in distribution to a normal distribution with mean zero where* $\widehat{\boldsymbol{\beta}}^1$ *and* $\boldsymbol{\beta}_0^1$ *denote the components of* $\widehat{\boldsymbol{\beta}}$ *and* $\boldsymbol{\beta}_0$ *corresponding to non-zero* β_{0k}*'s.*

Simulations

Simulation studies: simulation design

Key settings:

- $Z_i|D_i = d \sim N(\mu_d, \sigma_d^2)$
- $\mu_1 - \mu_{-1} = 1.5$ or 2, for continuous and multinomial $\boldsymbol{X}_i$, respectively
- $\boldsymbol{X}_i = (X_{i1}, \cdots, X_{i10})^T|D_i = d \sim MVN(\boldsymbol{\theta}_d, \boldsymbol{\Sigma}_d)$ for continuous $\boldsymbol{X}_i$, thresholding for binary, and multinomial

Methods to be compared:

- SVM-Oracle: assuming D_i is known
- SVM-2stage: Gaussian mixture model analysis with Z, use obtained labels as outcomes and $\boldsymbol{X}$ as features in an SVM
- SVM-EM with various penalty functions

Simulation studies: simulation design

Four scenarios:

I: $\boldsymbol{X}$ are continuous and independent features;

II: $\boldsymbol{X}$ are binary and independent features;

III: $\boldsymbol{X}$ are multinomial and independent features ($X_{ij} = 0, 1, \cdots, 4$);

IV: $\boldsymbol{X}$ are continuous and correlated features;

Simulation studies: results

Table: Prediction and feature selection performance (n=500)

Setting	Method	Miss	AUC	C	IC
I	SVM-Oracle	0.043	0.993	2.990	0.002
	SVM-EM	0.050	0.991	2.928	0.054
	SVM-2stage	0.075	0.987	2.948	0.046
II	SVM-Oracle	0.094	0.864	2.994	0.001
	SVM-EM	0.104	0.861	2.950	0.410
	SVM-2stage	0.130	0.814	2.618	0.010
III	SVM-Oracle	0.102	0.955	2.998	0.001
	SVM-EM	0.104	0.955	3.000	0.142
	SVM-2stage	0.115	0.953	2.988	0.472
IV	SVM-Oracle	0.045	0.992	2.990	0.162
	SVM-EM	0.054	0.989	2.900	0.316
	SVM-2stage	0.076	0.986	2.938	0.292
	SVM-Oracle*	0.044	0.992	3.000	0.066
	SVM-EM*	0.054	0.991	2.998	0.001
	SVM-2stage*	0.073	0.990	2.988	0.096

Simulation studies: sensitivity analysis

Four scenarios:

A: Misspecify the distribution of Z_i as normal while generated from Laplace;

B: Misspecify $\boldsymbol{X}$ and Z as conditionally independent given D_i while they are correlated;

C: Alternative pseudo-class probabilities:

$$\Pr^{(m)}[D_i = d_i^{(m)} | \boldsymbol{X}_i] = \frac{1}{1 + [1 - d_i^{(m)} g(\boldsymbol{X}_i)]_+}.$$

D: Randomly reveal 10%, 20% or 30% of disease labels and semi-supervised learning in literature.

Simulation studies: sensitivity analysis

Table: Prediction and feature selection performance (n=500)

Setting	Method	Miss	AUC	C	IC
A	SVM-Oracle	0.043	0.992	2.990	0.002
	SVM-EM	0.059	0.989	2.866	0.082
	SVM-2stage	0.083	0.987	2.876	0.078
B	SVM-Oracle	0.044	0.992	2.984	0.002
	SVM-EM	0.061	0.988	2.864	0.390
	SVM-2stage	0.074	0.987	2.946	0.526
C	SVM-Oracle	0.097	0.857	2.994	0.000
	SVM-EM	0.114	0.844	2.864	0.156
	SVM-2stage	0.132	0.809	2.612	0.002
D	Semi-10%	0.059	0.988	2.878	0.454
	Semi-20%	0.049	0.991	2.922	0.118
	Semi-30%	0.045	0.992	2.958	0.046
	SVM-EM	0.050	0.991	2.928	0.054

Applications

Application to CG studies

Training data:

- Pittsburgh study (Shear et al. 2005): CG psychotherapy (CGT) compared to interpersonal psychotherapy (IPT)
- 175 subjects (67% women) with a baseline visit

Independent validation data:

- HEAL (on going): multi-site study to compare response to antidepressant medication administered with and without CGT among bereaved individuals
- CGTOA (on going): CGT in an older population (≥ 50)
- 196 subjects (109 HEAL, 87 CGTOA)

Fit linear decision rule for easy interpretation.

Application to CG studies

Assessments of CG symptoms ($\boldsymbol{X}$):

- Inventory of Complicated Grief (ICG) (Prigerson et al. 1995).

Informative marker (Z):

- Work and Social Adjustment Scale (WSAS) (Mundt et al. 2002), range 0-40.

Validation variables:

- Structured Clinical Interview of Complicated Grief (SCI-CG)
- Impairment in social suport (ISEL)
- Depression score (HAM-D)
- Impact of event scales (IES-T, IES-I, IES-A)

Application to CG studies: results

Indirectly assessing validity:

Table: Correlation with other external measures on independent data sets: HEAL and CGTOA.

Measure	SVM1*	SVM2†	Z as GS‡
SCI-CG	0.47	0.38	0.34
Ham-D	0.58	0.50	0.57
ISEL	0.43	0.35	0.38
IES-T	0.26	0.19	0.18
IES-I	0.22	0.19	0.19
IES-A	0.24	0.15	0.13

*: SVM-EM with Enet penalty

†: SVM-EM without variable selection

‡: WSAS as the outcome for penalized regression with elastic net penalty

Application to CG studies: results

Table: Selected variables from the ICG and their coefficients in the Pittsburgh complicated grief study (training data).

Domain 1: Yearning and preoccupation with the deceased (5 items)	
I think about this person so much	
that it's hard for me to do the things I normally do	0.22
Life is empty without the person who died	0.09
Unfair that I should live when this person died	0.10
Feel lonely a great deal of the time ever since death	0.12
Domain 2: Anger and bitterness (2 items)	
I feel bitter over this persons death	0.01
Domain 3: Shock and disbelief (3 items)	
Disbelief over what happened	0.02
Domain 4: Estrangement from others (3 items)	
Lost the ability to care about other people	0.13
Envious of others who have not lost someone close	0.04
Domain 5: Hallucinations of the deceased (3 items)	
	None
Domain 6: Behavior change, including avoidance or proximity seeking (3 items)	
	None

Comparing to existing literature

Simon et al. (2011): clinically confirmed CG and non-CG

	Prevalence in cases (n=288)	Prevalence in non-cases (n=377)	
	Sensitivity	False positive	Specificity
Symptom Clusters			
1. Yearning and preoccupation with the deceased (5 items)	96.9%	11.7%	88.3%
2. Anger and bitterness (2 items)	72.6%	5.3%	94.7%
3. Shock and disbelief (3 items)	87.2%	6.1%	93.9%
4. Estrangement from others (3 items)	76.7%	5.8%	94.2%
5. Hallucinations of the deceased (3 items)	24.0%	1.6%	98.4%
6. Behavior change, including avoidance or proximity seeking (3 items)	92.4%	13.8%	86.2%

Comparing to existing literature

Simon et al. (2011): clinically confirmed CG and non-CG

TABLE 5. Sensitivity and specificity of ICG items in complicated grief cases versus noncases

	Prevalence in cases ($n = 288$)	Prevalence in noncases ($n = 377$)		
	Sensitivity (true-positive rate) (%)	False-positive rate (%)	Specificity (1−false-positive rate) (%)	Ratio of test positive
4. I feel myself longing for the person who died	88.5	9.3	90.7	9.5
2. Memories of the person who died upset me	81.9	7.2	92.8	11.4
19. I feel lonely a great deal of the time ever since he/she died	80.9	2.4	97.6	33.9
13. I feel that life is empty without the person who died	80.2	1.9	98.1	43.2
7. I feel disbelief over what happened	76.4	3.7	96.3	20.6
8. I feel stunned or dazed over what happened	71.2	1.9	98.1	38.3
3. I feel I cannot accept the death of the person who died	70.1	1.9	98.1	37.8
6. I can't help feeling angry about his/her death	64.6	3.7	96.3	17.4
1. I think about this person so much, it's hard to do the things I normally do	61.1	1.3	98.7	46.1
17. I feel bitter over this person's death	61.1	2.1	97.9	28.8
9. Ever since he/she died it is hard for me to trust people	49.0	2.4	97.6	20.5
10. Ever since he/she died I have lost the ability to care about other people	47.9	1.6	98.4	30.1
5. I feel drawn to places and things associated with the person who died	46.5	5.3	94.7	8.8
18. I feel envious of others who have not lost someone close	45.1	2.7	97.3	17.0
16. I feel that it is unfair that I should live when this person died	40.6	0.0	100.0	[a]9,999
12. I go out of my way to avoid reminders of the person who died	32.3	2.4	97.6	13.5
11. I have pain in the same area of my body as the person who died	12.8	1.3	98.7	9.7
15. I see the person who died stand before me	10.8	0.3	99.7	40.6
14. I hear the voice of the person who died speak to me	9.7	0.0	100.0	9,999[a]

[a]Values of 9,999 are given for those cases where the ratio is estimated to be infinite (due to a zero false-positive rate in the noncase group). Likelihood ratio of test positive = sensitivity/(1−specificity).

UCI data analysis

UCI machine learning repository:

- Wisconsin breast cancer (WBC)
- Pima Indians diabetes (PIMA)
- HEART
- Spam email (SPAM)

Class labels known, but not used.

Combine 20% to 30% of the attributes as informative markers.

SPAM: 1000 randomly selected as training, the rest testing.

Three others: 200 cases were randomly selected as training.

UCI data analysis results

Table: Analysis results of four UCI Machine Learning Repository data sets

		WBC	SPAM	PIMA	HEART
Number of feature variables		7	47	7	9
Correlation†		0.867	0.695	0.443	0.378
Miss	SVM-EM	0.045	0.161	0.365	0.267
	SVM-Oracle	0.043	0.130	0.325	0.197
AUC	SVM-EM	0.986	0.889	0.643	0.806
	SVM-Oracle	0.988	0.931	0.703	0.866

†: Pearson correlation between the informative marker and disease status

Discussion

Discussion

- Incorporating biomarkers (e.g., polysomnography for sleep-wake disorders)
- Multi-class problems
- Joint learning to relax conditional independence assumption
- Multivariate informative markers
- Semi-supervised learning
- Other pseudo class probabilities
- Incorporating correlation structure