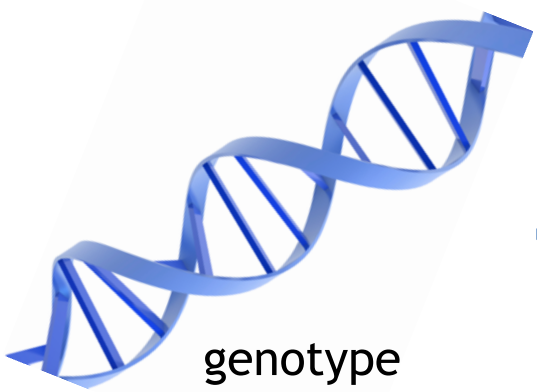
Separating sparse signals from correlated noise in binary classification

Stephan Mandt
Columbia University

Florian Wenzel, Shinichi Nakajima, Christoph Lippert, Marius Kloft

Statistical genetics: predict the effect of genes on observable outcomes



genotype

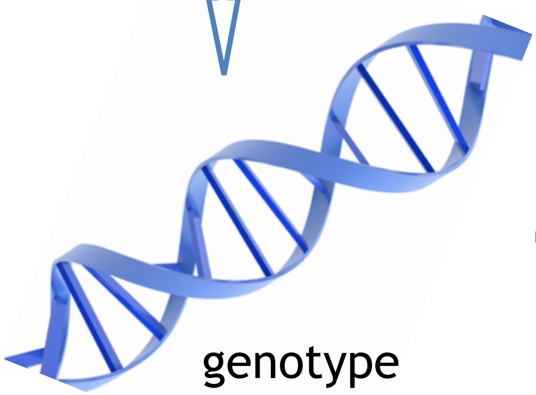
causal association $\Longrightarrow$

$$\{-1, +1\}$$

binary phenotype
(e.g. a disease)

Statistical genetics: predict the effect of genes on observable outcomes

genotypes correlate with
geographic location



genotype

causal association

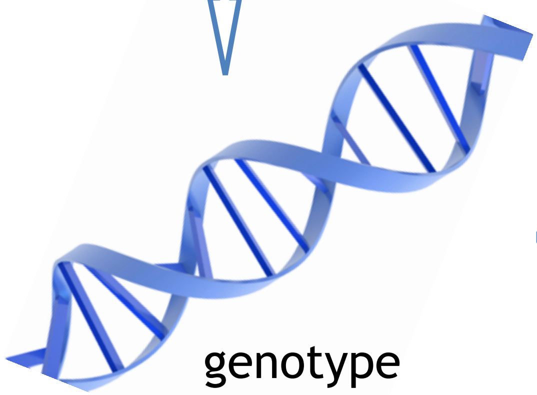
$$\{-1, +1\}$$

binary phenotype
(e.g. a disease)

Statistical genetics: predict the effect of genes on observable outcomes

genotypes correlate with
geographic location

cultural factors
climate factors
economic conditions
political factors



genotype

causal association

$$\{-1, +1\}$$

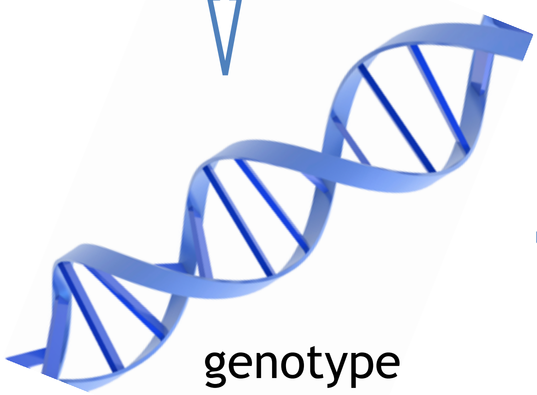
binary phenotype
(e.g. a disease)

Statistical genetics: predict the effect of genes on observable outcomes

genotypes correlate with geographic location

Problem:
indirect associations
"population structure"
Goal:
separate these effects

cultural factors
climate factors
economic conditions
political factors



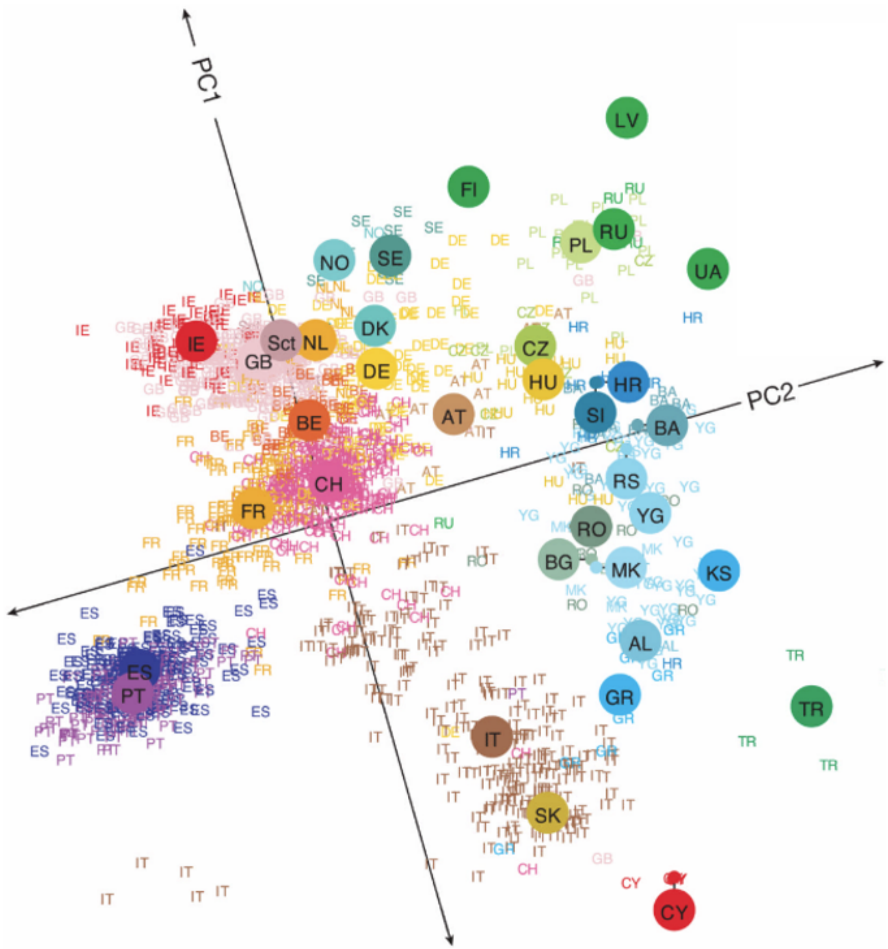
genotype

causal association

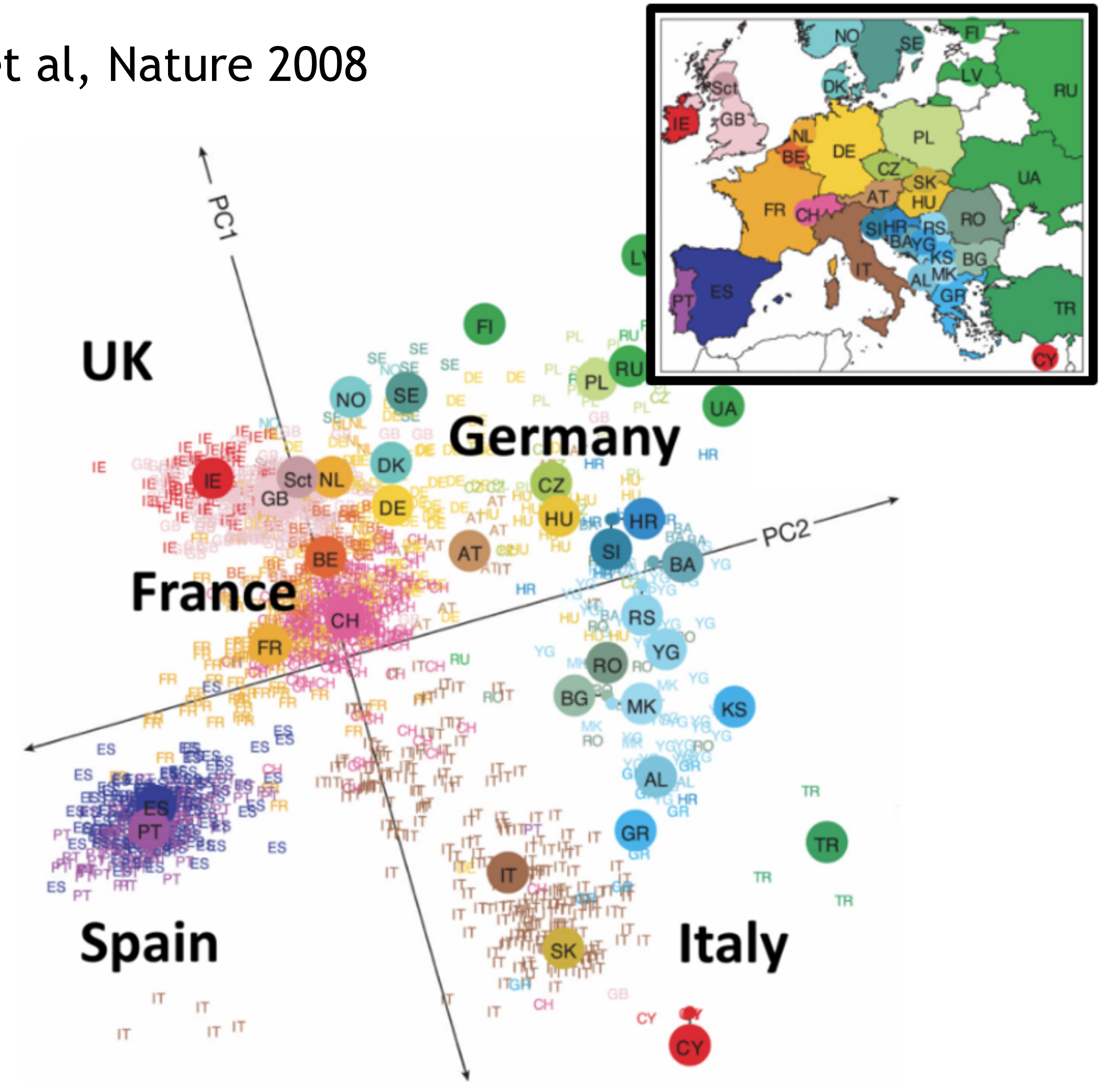
$$\{-1, +1\}$$

binary phenotype
(e.g. a disease)

Novembre et al, Nature 2008

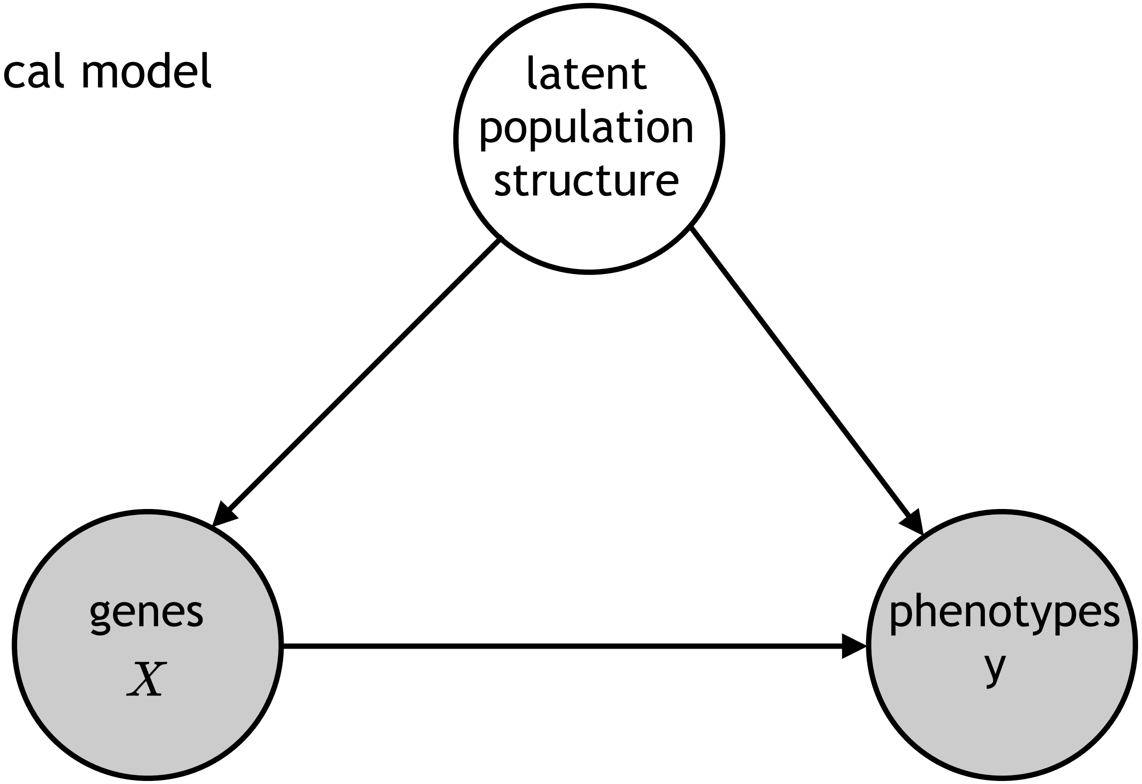


Novembre et al, Nature 2008



Confounding by population structure

graphical model



- Population structure: common cause of genes and phenotypes
- Goal: estimate causal effect
- Problem: generative mechanism is unknown

Linear Mixed Models for Regression

- Linear mixed model (LMM): widely appreciated in genetics
- Linear regression + correlated noise.

$$
\begin{aligned}
X \in \mathbb{R}^{d\times n} &: \text{genotypes} \\
y \in \mathbb{R}^{n} &: \text{phenotypes} \\
w \in \mathbb{R}^{d} &: \text{weight vector}
\end{aligned}
$$

$$y_i = X_i^\top w + \epsilon_i \qquad \epsilon \sim \mathcal{N}(0, \Sigma)$$

Lippert, Christoph, et al. "FaST linear mixed models for genome-wide association studies." Nature Methods 8.10 (2011): 833-835.
Rakitsch, Barbara, et al. "A Lasso multi-marker mixed model for association mapping with population structure correction." Bioinformatics 29.2 (2013): 206-214.

Linear Mixed Models for Regression

- Linear mixed model (LMM): widely appreciated in genetics
- Linear regression + correlated noise.

$$X \in \mathbb{R}^{d\times n} \;:\; \text{genotypes}$$
$$y \in \mathbb{R}^{n} \;:\; \text{phenotypes}$$
$$w \in \mathbb{R}^{d} \;:\; \text{weight vector}$$

$$y_i = X_i^\top w + \epsilon_i \qquad \epsilon \sim \mathcal{N}(0, \Sigma)$$

Multivariate noise:

- Allows to express similarities between samples
- Typical choice in genetics: $\Sigma = \lambda_0 \mathbf{I} + \lambda_1 X^\top X$

Lippert, Christoph, et al. "FaST linear mixed models for genome-wide association studies." Nature Methods 8.10 (2011): 833-835.
Rakitsch, Barbara, et al. "A Lasso multi-marker mixed model for association mapping with population structure correction." Bioinformatics 29.2 (2013): 206-214.

LMMs: why using a linear kernel?

Equivalent model formulation with a linear kernel:

$$y = X^\top w + X^\top w' + \epsilon, \quad w' \sim \mathcal{N}(0, \lambda_1 \mathbf{I}), \quad \epsilon \sim \mathcal{N}(0, \lambda_0 \mathbf{I})$$

We see that *w* and *w'* play similar roles, **but:**

- *w* is assumed to be a large, **sparse** vector (causal)
- w' is dense and unobserved (confounder)

Lippert, Christoph, et al. "FaST linear mixed models for genome-wide association studies." Nature Methods 8.10 (2011): 833-835.
Rakitsch, Barbara, et al. "A Lasso multi-marker mixed model for association mapping with population structure correction." Bioinformatics 29.2 (2013): 206-214.

Solving linear mixed models

A solution can be obtained by transforming *X* and *y*:

$$\Sigma = BB^\top, \quad y' = B^{-1}y, \quad X'^\top = B^{-1}X^\top$$

$$y' = X'^\top w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

- This results in a standard linear regression problem
- O(n^3) scaling if done naively
- State of the art for many applications in biology

Lippert, Christoph, et al. "FaST linear mixed models for genome-wide association studies." Nature Methods 8.10 (2011): 833-835.
Rakitsch, Barbara, et al. "A Lasso multi-marker mixed model for association mapping with population structure correction." Bioinformatics 29.2 (2013): 206-214.

This talk: linear mixed model for binary classification

Goal: generalize the LMM paradigm to classification

Idea: probit regression model with correlated noise:

$$y_i = \mathrm{sign}(X_i^\top w + \epsilon_i), \quad \epsilon \sim \mathcal{N}(0, \Sigma)$$

Challenge: exact inference becomes intractable due to the nonlinearity

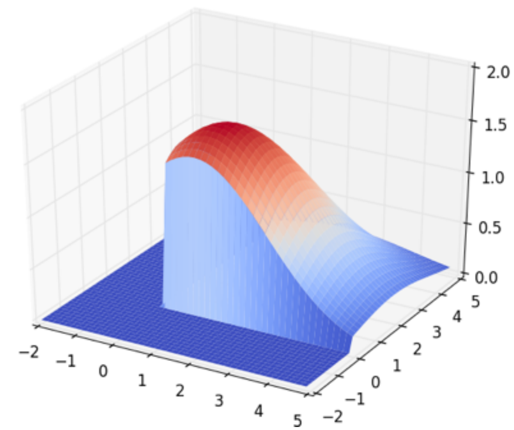
Solution: approximate inference (this talk)

Deriving a loss function

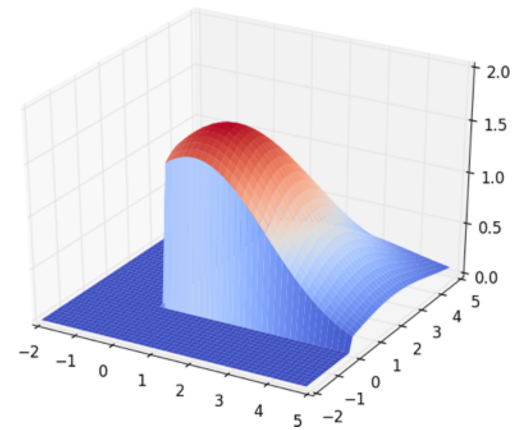
For simplicity, assume $\forall_i : y_i = 1$

Likelihood that all examples are correctly classified:

$$\mathbb{P}\big(\forall_i : y_i = \text{sign}(X_i^\top w + \epsilon_i)\big) = \int_{\mathbb{R}_+^n} \mathcal{N}(\epsilon; X^\top w, \Sigma) d^n \epsilon$$



Deriving a loss function



For simplicity, assume $\forall_i : y_i = 1$

Likelihood that all examples are correctly classified:

$$\mathbb{P}\big(\forall_i : y_i = \mathrm{sign}(X_i^\top w + \epsilon_i)\big) = \int_{\mathbb{R}_+^n} \mathcal{N}(\epsilon; X^\top w, \Sigma) d^n \epsilon$$

Objective function: negative log likelihood + regularizer.

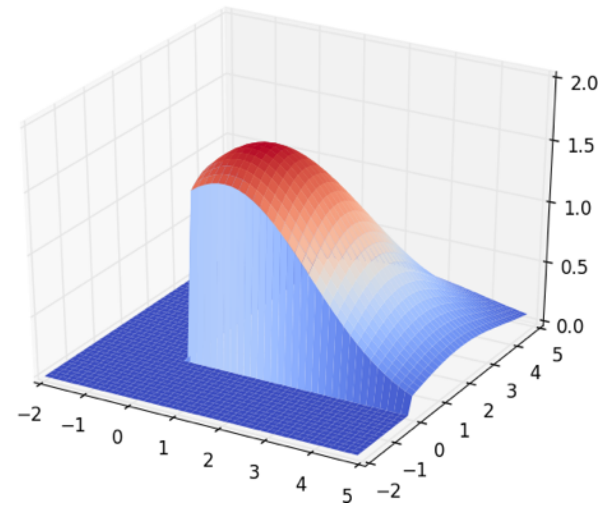
Define $\mu(w) = X^\top w$.

$$\boxed{\mathcal{L}(w) = \underbrace{-\log \int_{\mathbb{R}_+^n} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\mathrm{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\mathrm{reg}}(w)}}$$

Lasso regularizer:
Favors sparsity

Minimizing the loss function

$$\mathcal{L}(w) \;=\; \underbrace{-\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\text{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\text{reg}}(w)}$$



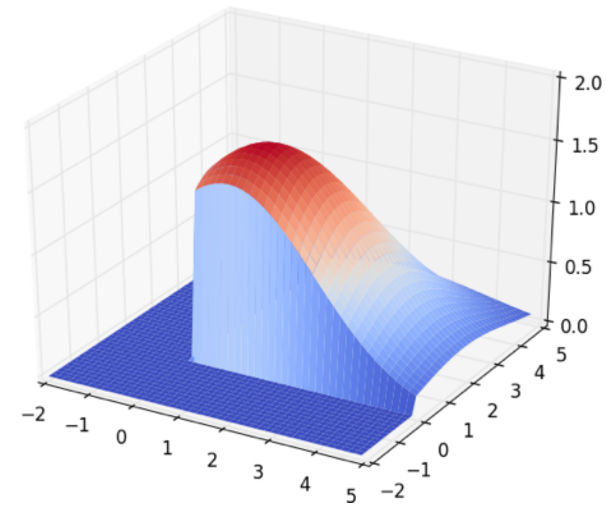
Minimizing the objective function leads to two computational problems:

(i) intractable high-dimensional integral

(ii) the l_1-norm regularizer is not everywhere differentiable

Minimizing the loss function

$$\mathcal{L}(w) = \underbrace{-\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\text{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\text{reg}}(w)}$$



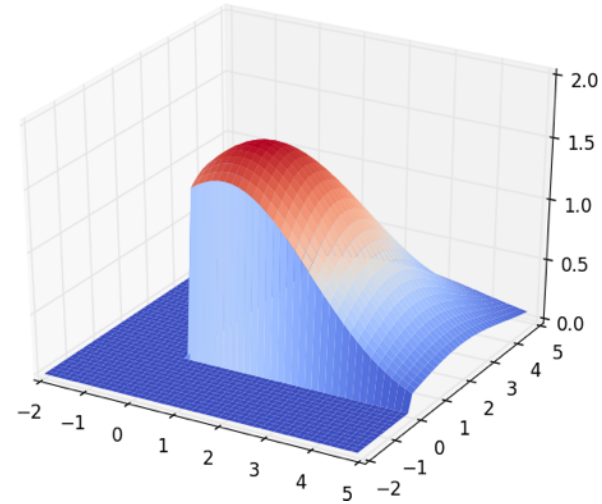
Minimizing the objective function leads to two computational problems:

(i) intractable high-dimensional integral
Solution: Expectation Propagation (EP)

(ii) the l_1-norm regularizer is not everywhere differentiable

Minimizing the loss function

$$\mathcal{L}(w) \;=\; \underbrace{-\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\text{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\text{reg}}(w)}$$



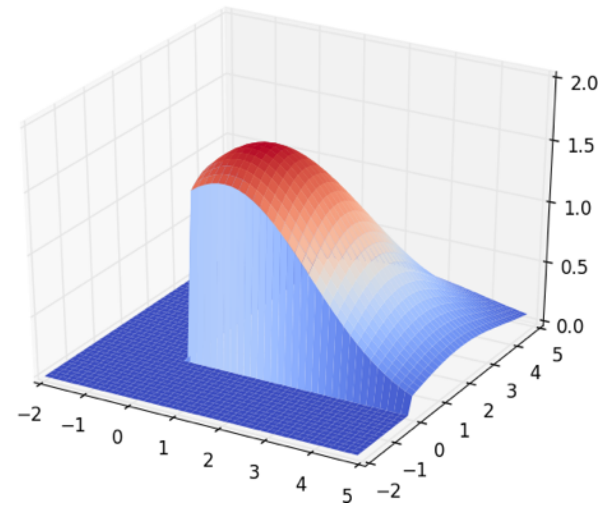
Minimizing the objective function leads to two computational problems:

(i) intractable high-dimensional integral
Solution: Expectation Propagation (EP)

(ii) the l_1-norm regularizer is not everywhere differentiable
Solution: Alternating Direction Method of Multipliers (ADMM)

Minimizing the Likelihood Term

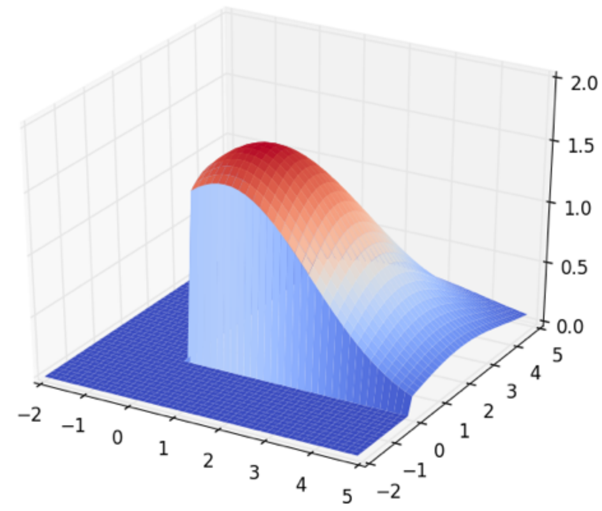
$$\mathcal{L}(w) = \underbrace{-\log \int_{\mathbb{R}_+^n} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\mathrm{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\mathrm{reg}}(w)}$$



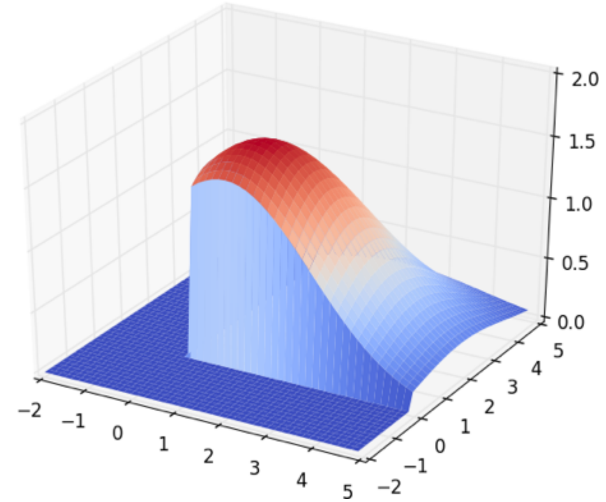
Minimizing the Likelihood Term

$$\mathcal{L}(w) \;=\; \underbrace{-\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\text{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\text{reg}}(w)}$$

$$p(\epsilon|\mu, \Sigma) \;=\; \frac{\mathbb{1}[\epsilon \in \mathbb{R}^n_+]\, \mathcal{N}(\epsilon; \mu, \Sigma)}{\int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu, \Sigma)\, d^n\epsilon}$$



Minimizing the Likelihood Term



$$\mathcal{L}(w) \quad = \quad \underbrace{-\log \int_{\mathbb{R}_+^n} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\text{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\text{reg}}(w)}$$

$$p(\epsilon|\mu, \Sigma) \quad = \quad \frac{\mathbb{1}[\epsilon \in \mathbb{R}_+^n]\, \mathcal{N}(\epsilon; \mu, \Sigma)}{\int_{\mathbb{R}_+^n} \mathcal{N}(\epsilon; \mu, \Sigma)\, d^n\epsilon}$$

$$\mu_p(w) \quad = \quad \mathbb{E}_{p(\epsilon|\mu(w),\Sigma)}\left[\epsilon\right],$$

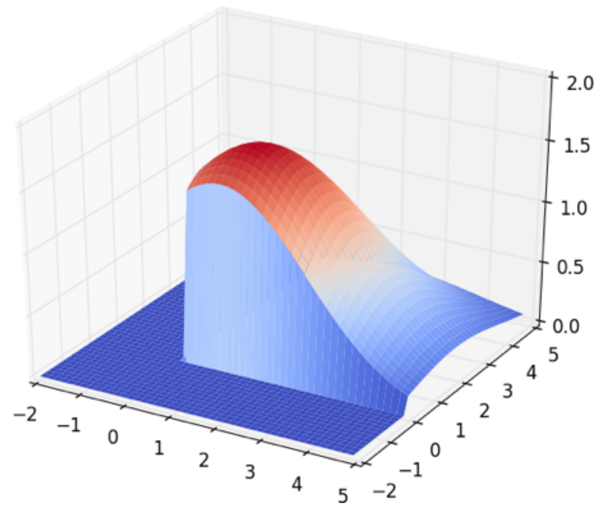
$$\Sigma_p(w) \quad = \quad \mathbb{E}_{p(\epsilon|\mu(w),\Sigma)}\left[(\epsilon - \mu_p(w))(\epsilon - \mu_p(w))^\top\right]$$

$$\Delta\mu = \mu_p - \mu$$

$$\boxed{\begin{aligned} \nabla_w \mathcal{L}^{\text{loss}}(w) \quad &= \quad \Delta\mu \Sigma^{-1} X^\top, \\ H^{\text{loss}}(w) \quad &= \quad -X[\Sigma^{-1}(\Sigma_p - \Delta\mu\Delta\mu^\top)\Sigma^{-1} - \Sigma^{-1}]X^\top \end{aligned}}$$

Minimizing the Likelihood Term

$$\mathcal{L}(w) = \underbrace{-\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\text{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\text{reg}}(w)}$$



$$p(\epsilon|\mu, \Sigma) = \frac{\mathbb{1}[\epsilon \in \mathbb{R}^n_+]\, \mathcal{N}(\epsilon; \mu, \Sigma)}{\int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu, \Sigma)\, d^n\epsilon}$$

$$\mu_p(w) = \mathbb{E}_{p(\epsilon|\mu(w),\Sigma)}[\epsilon],$$

$$\Sigma_p(w) = \mathbb{E}_{p(\epsilon|\mu(w),\Sigma)}\left[(\epsilon - \mu_p(w))(\epsilon - \mu_p(w))^\top\right]$$

$$\Delta\mu = \mu_p - \mu$$

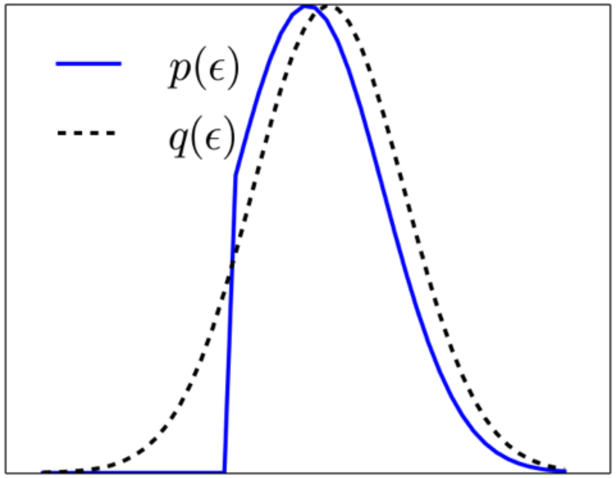
Still intractable integrals. Need approximate inference

$$\boxed{\begin{aligned} \nabla_w \mathcal{L}^{\text{loss}}(w) &= \Delta\mu \Sigma^{-1} X^\top, \\ H^{\text{loss}}(w) &= -X[\Sigma^{-1}(\Sigma_p - \Delta\mu\Delta\mu^\top)\Sigma^{-1} - \Sigma^{-1}]X^\top \end{aligned}}$$

Minimizing the Likelihood Term

$$\mathcal{L}(w) = \underbrace{-\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu(w), \Sigma)\, d^n\epsilon}_{=:\mathcal{L}^{\text{loss}}(w)} + \underbrace{\lambda_0 ||w||_1^1}_{=:\mathcal{L}^{\text{reg}}(w)}$$

$$p(\epsilon|\mu, \Sigma) = \frac{\mathbb{1}[\epsilon \in \mathbb{R}^n_+]\, \mathcal{N}(\epsilon; \mu, \Sigma)}{\int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; \mu, \Sigma)\, d^n\epsilon}$$



[J. Cunningham et. al., Gaussian probabilities and EP, arxiv 2011.]

We use **Expectation Propagation** to approximate $p(\epsilon|\mu, \Sigma)$ by a Gaussian $q(\epsilon; \mu_q, \Sigma_q) = \mathcal{N}(\epsilon; \mu_q, \Sigma_q)$

Then: $\mu_p \approx \mu_q$ $\Sigma_p \approx \Sigma_q$

Comparison to other methods

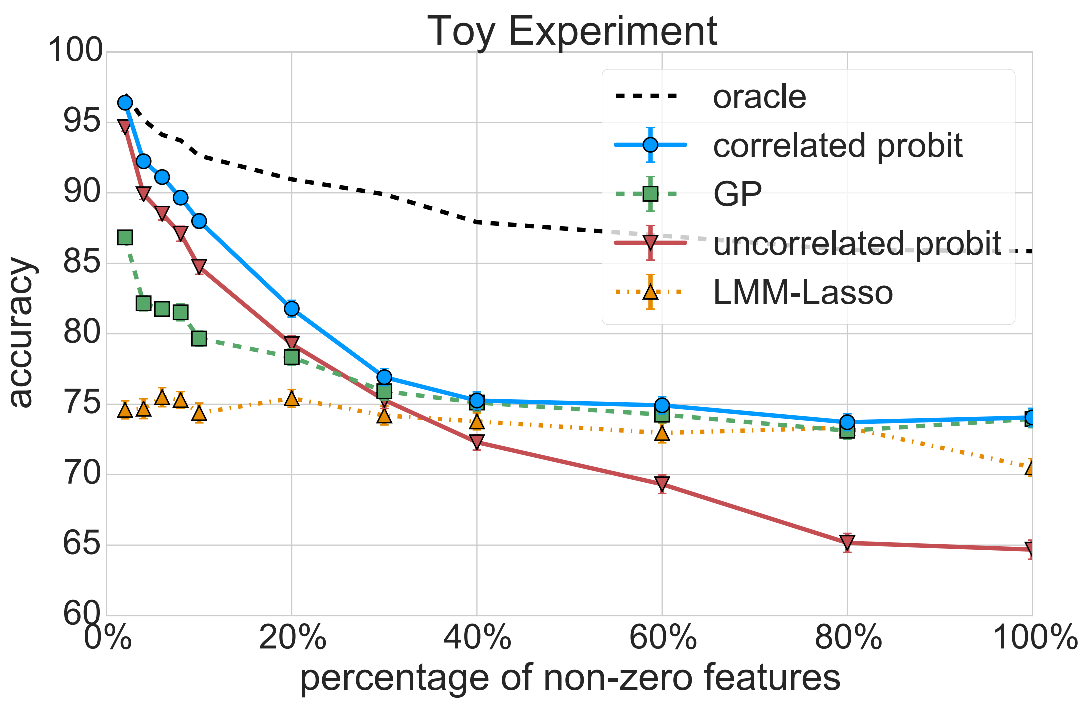
- The correlated probit model can be seen as a generalization of various other models.
- We compared the performance gain over these methods in our experiments.

Uncorrelated Probit: $y = \mathrm{sign}(X^\top w + \epsilon), \quad \epsilon \sim \mathcal{N}(0, \mathrm{I}), \quad w \sim \mathrm{Laplace}(\,.\,; \lambda_0)$

GP classification: $y = \mathrm{sign}(f), \quad f \sim \mathcal{N}(0, \Sigma(X))$

LMM Lasso: $y = X^\top w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \Sigma), \quad w \sim \mathrm{Laplace}(\,.\,; \lambda_0)$

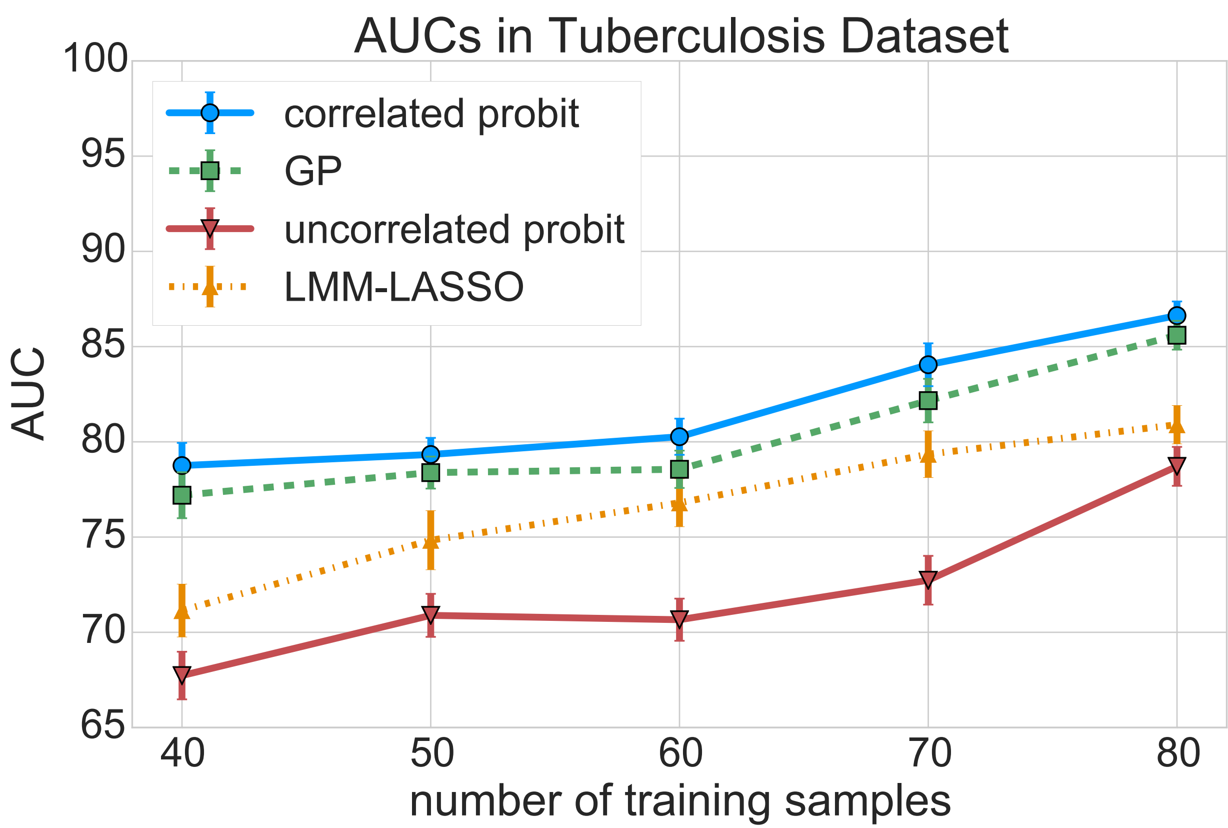
- Generate artificial data from the model
- Varied the amount of non-zero weights w_i
- Compute accuracies for different levels of sparsity



Experiments (TBC)

- **Predict Tuberculosis** based on gene expression levels.
- **Confounding** by populations structure.

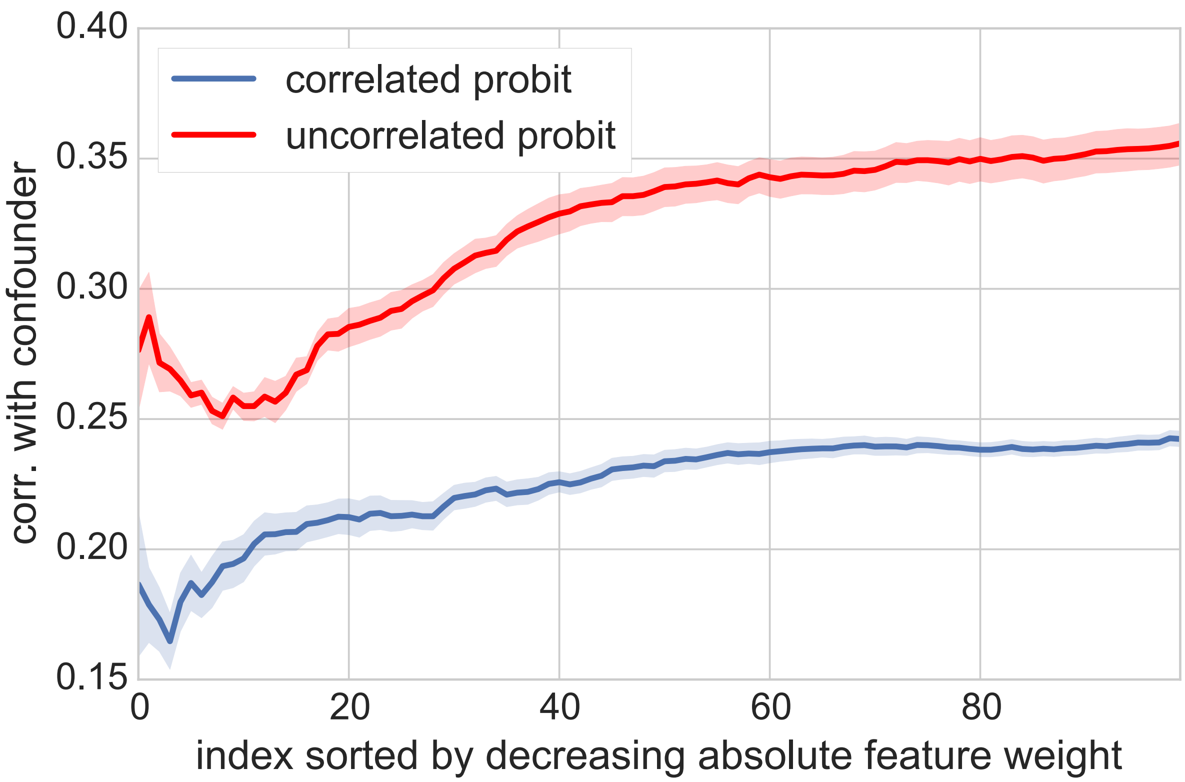
Tuberculosis data set:
Berry et. al., Nature 466, 2010.



Experiments (TBC)

- **Tuberculosis** data set
- Correlate w with largest eigenvalue of Σ

Tuberculosis data set:
Berry et. al., Nature 466, 2010.

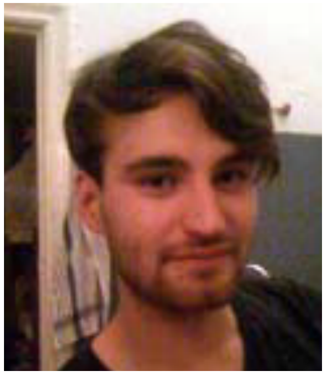


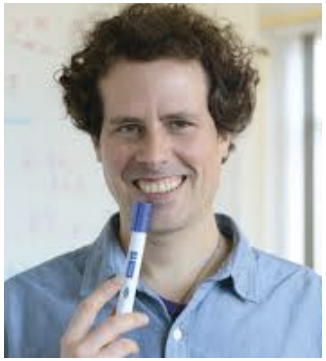
Conclusion and Outlook

- Algorithm for sparse feature selection in binary classification, where the data are confounded
- Signals found by our model are less correlated with the confounders
- Improved prediction performances

- Future: employ scalable MCMC to sample from the posterior
- Data subsampling is possible
- problem: high-dimensional feature space dimensionality d.

Thank you.



Florian Wenzel*
HU Berlin

Shinichi Nakajima
TU Berlin

Christoph Lippert
Human Longevity Inc.

Marius Kloft
HU Berlin

Links to publications: www.stephanmandt.com

* equal contribution