Correction of Confounding in Genome-Wide Association Studies

Florian Wenzel (HU Berlin)

This is joint work with:
Stephan Mandt (Disney Research)
Shinichi Nakajima (TU Berlin)
John Cunningham (Columbia University)
Christoph Lippert (Human Longevity)
Marius Kloft (HU Berlin)

Genome-Wide Association Studies (GWAS)

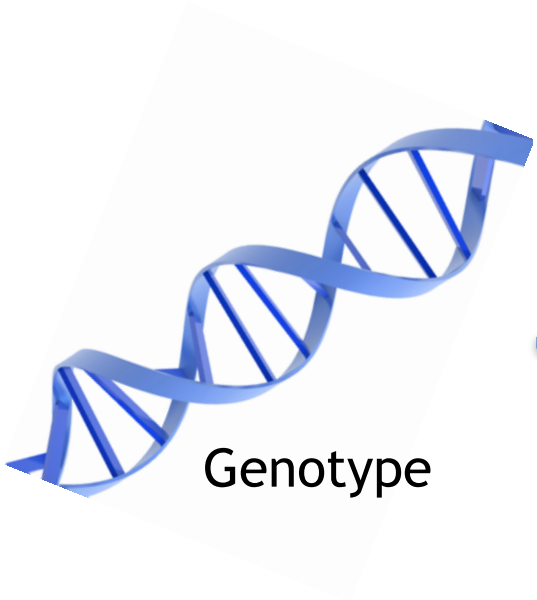
- Dataset of genetic variants in different individuals.
- Goal: find variants which are associated with a trait (e.g. a disease).

What can go wrong in GWAS?

Confounders can lead to spurious correlations between genotype and phenotype.

⇨ False Associations

Genetic Associations and Confounding



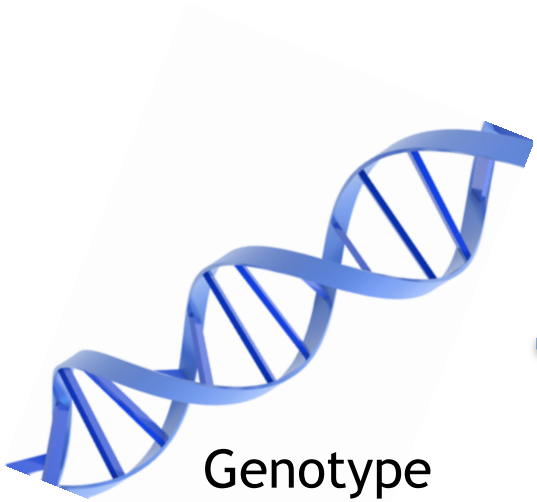
Genotype

Genetic association

Binary Phenotype
(e.g. a disease)

Genetic Associations and Confounding

- Economic conditions
- Cultural habits (eating/drinking)
- Geographical factors etc.



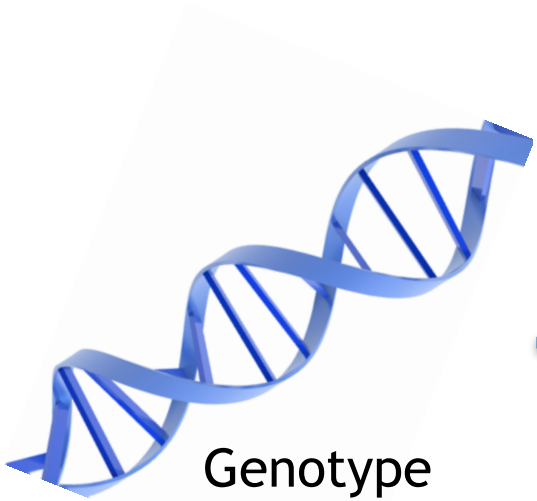
Genotype

Genetic association

Binary Phenotype
(e.g. a disease)

Genetic Associations and Confounding

- Economic conditions
- Cultural habits (eating/drinking)
- Geographical factors etc.



Genotype

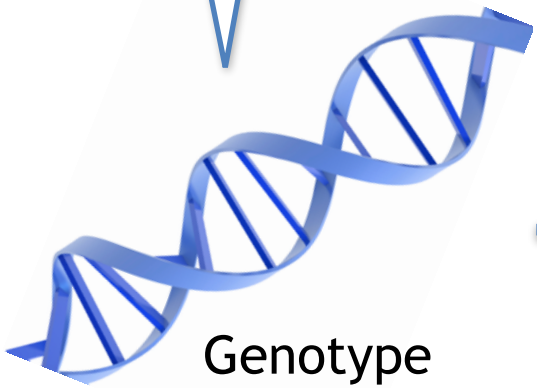
Genetic association

Binary Phenotype
(e.g. a disease)

Genetic Associations and Confounding

- Genetic similarity with the group

- Economic conditions
- Cultural habits (eating/drinking)
- Geographical factors etc.



Genotype

Genetic association

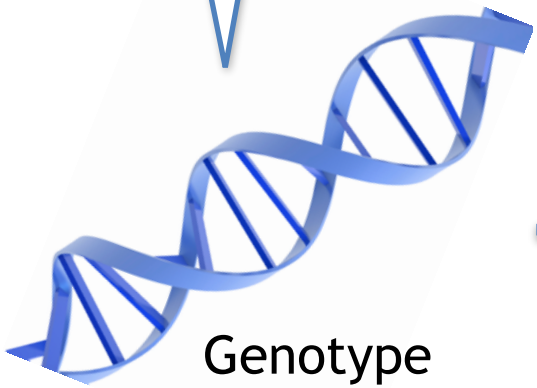
Binary Phenotype
(e.g. a disease)

Genetic Associations and Confounding

Population structure

- Genetic similarity with the group

- Economic conditions
- Cultural habits (eating/drinking)
- Geographical factors etc.

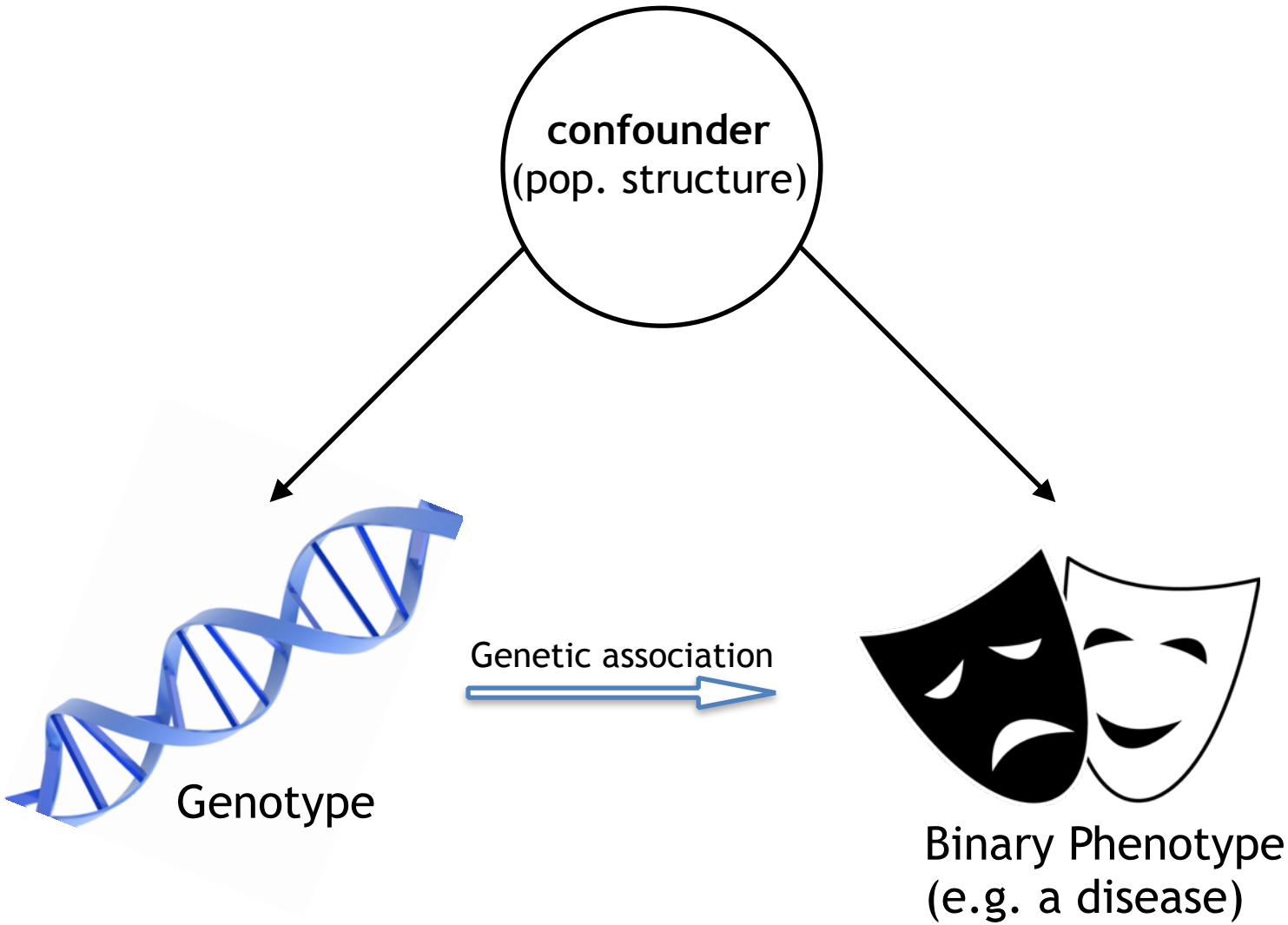


Genotype

Genetic association

Binary Phenotype
(e.g. a disease)

Genetic Associations and Confounding



Research Goal

We want a method that:

- **predicts a binary phenotype** based on the genome.
- finds genes which could be **associated with a binary phenotype** (e.g. disease yes/no).
- accounts for **confounding effects**.

Research Goal

We want a method that:

- **predicts a binary phenotype** based on the genome.
- finds genes which could be **associated with a binary phenotype** (e.g. disease yes/no).
- accounts for **confounding effects**.

We extend LMM-Lasso to the classification scenario.

New Method:

Correlated Probit Regression

LMM (Linear Mixed Model)

- Confounding effects can be modelled in terms of **correlated noise.**

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

X_i : d-dim data vector
y_i : label
w : d-dim parameter vector
ϵ_i : label noise

- The covariance matrix models similarities between subjects.
- Can be estimated from data (later more).

LMM-Lasso

- We want to have sparse representation.
- i.e. assigning zero-effect size to the majority of SNPs.
- Shrinkage prior over ***w***.

$$y_i = X_i^\top w + \epsilon_i \qquad \epsilon \sim \mathcal{N}(0, \Sigma)$$
$$w \sim \text{Laplace}(0, \lambda_0^{-1})$$

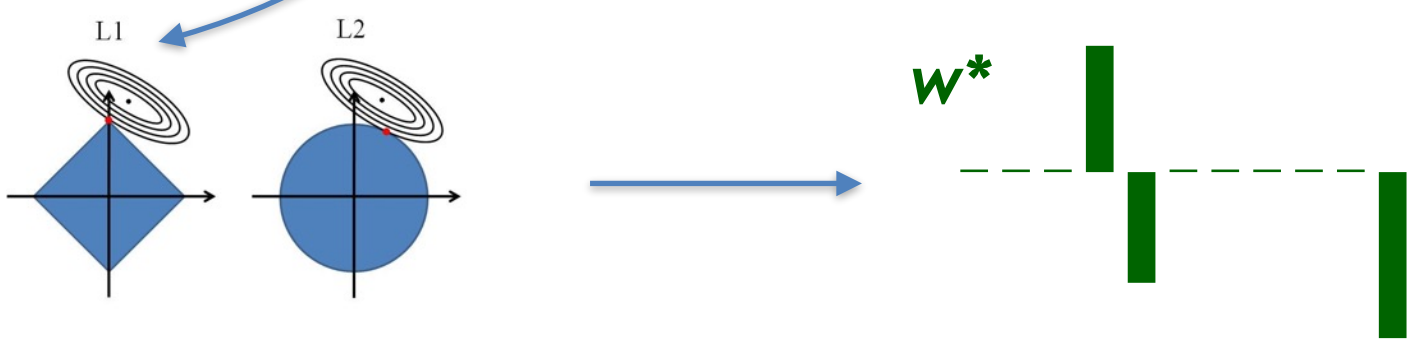
$$w^* = \arg\max_w \, p(w|X, y)$$

- **Inference: Compute MAP**

Rakitsch et al. „A Lasso Multi-Marker Mixed Model for Association Mapping with Population Structure Correction.“ Bioinformatics (2012).

- We want to have sparse representation.
- i.e. assigning zero-effect size to the majority of SNPs.
- Shrinkage prior over ***w***.

$$y_i = X_i^\top w + \epsilon_i \qquad \epsilon \sim \mathcal{N}(0, \Sigma)$$
$$w \sim \text{Laplace}(0, \lambda_0^{-1})$$



LMM-Lasso

LMM-Lasso very successful in genetic application.

Rakitsch et al. „A Lasso Multi-Marker Mixed Model for Association Mapping with Population Structure Correction.“ Bioinformatics (2012).

Lippert et al. „FaST linear mixed models for genome-wide association studies.“ Nature Methods 8.10 (2011): 833-835.

LMM-Lasso

LMM-Lasso very successful in genetic application.

Limitation: Only applicable for Regression!

Rakitsch et al. „A Lasso Multi-Marker Mixed Model for Association Mapping with Population Structure Correction.“ Bioinformatics (2012).

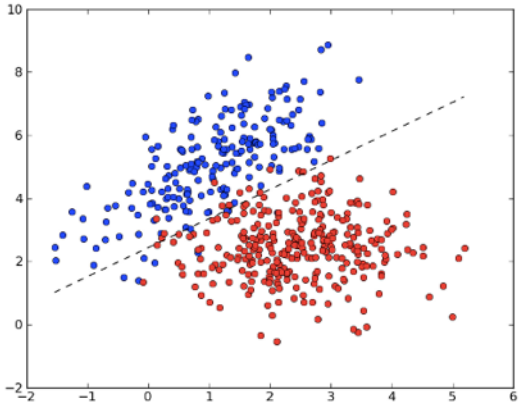
Lippert et al. „FaST linear mixed models for genome-wide association studies.“ Nature Methods 8.10 (2011): 833-835.

This talk: Correlated Probit Model

Extend this model for classification!

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

$$w \sim \text{Laplace}(0, \lambda_0^{-1})$$

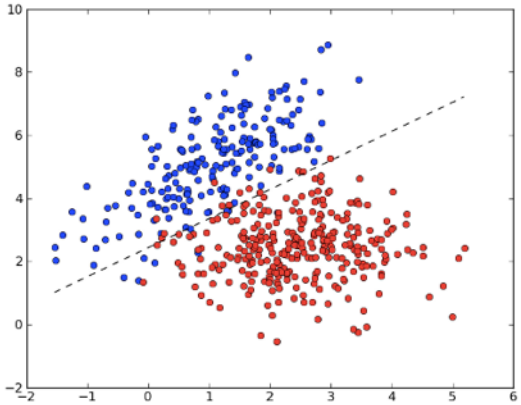


This talk: Correlated Probit Model

Extend this model for classification!

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

$$w \sim \mathrm{Laplace}(0, \lambda_0^{-1})$$



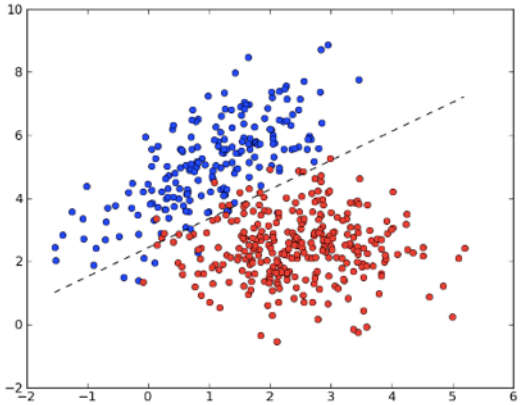
This talk: Correlated Probit Model

Correlated Probit model:

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

$$w \sim \mathrm{Laplace}(0, \lambda_0^{-1})$$

X : $d \times n$ data matrix
y : n-dim label vector
w : d-dim parameter vector
ϵ : n-dim noise vector



The Choice of the Covariance Matrix

$$\Sigma = \lambda_1 \mathbf{I} + \lambda_2 X^\top X + \lambda_3 \Sigma_{\text{side}}$$

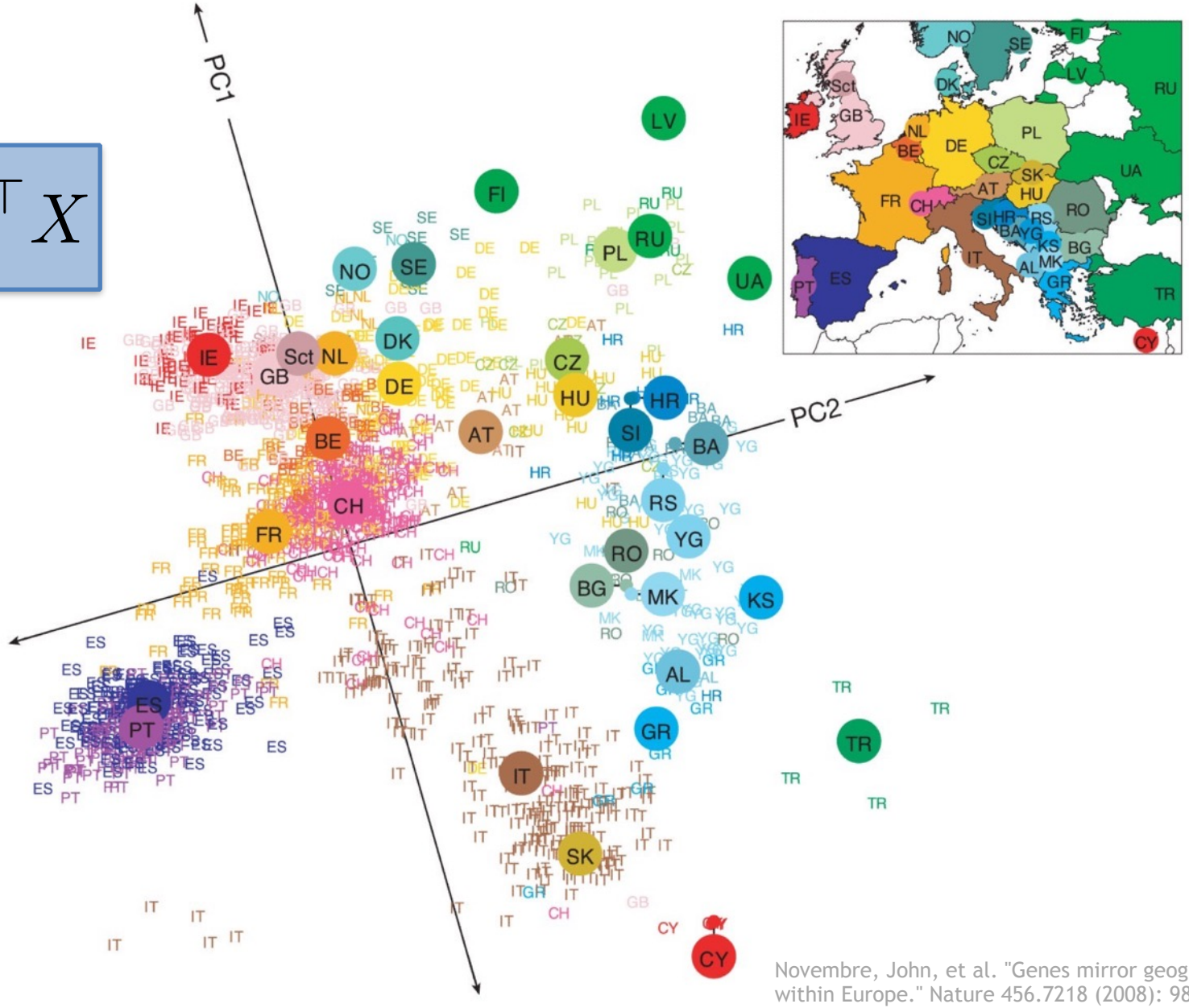
Measurement Noise

The Choice of the Covariance Matrix

$$\Sigma = \lambda_1 \mathbf{I} + \lambda_2 \boxed{X^\top X} + \lambda_3 \Sigma_{\text{side}}$$

Population Structure

$$X^\top X$$



Novembre, John, et al. "Genes mirror geography within Europe." Nature 456.7218 (2008): 98-101.

The Choice of the Covariance Matrix

$$\Sigma = \lambda_1 \mathbf{I} + \lambda_2 X^\top X + \lambda_3 \boxed{\Sigma_{\text{side}}}$$

Kernel on top of side information

The Choice of the Covariance Matrix

$$\Sigma = \lambda_1 \mathbf{I} + \lambda_2 X^\top X + \lambda_3 \Sigma_{\text{side}}$$

Weights are learned via cross validation.

Correlated Probit Regression

Correlated Probit model:

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

$$w \sim \mathrm{Laplace}(0, \lambda_0^{-1})$$

X : $d \times n$ data matrix
y : n-dim label vector
w : d-dim parameter vector
ϵ : n-dim noise vector

- Applies to classification.
- Captures correlations between the labels.
- Inference becomes harder.

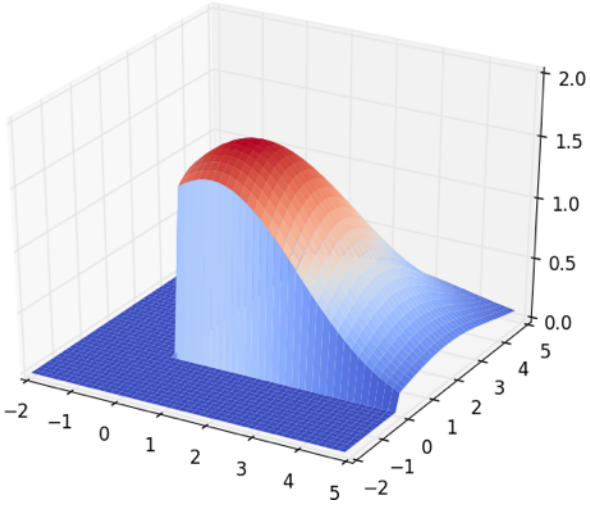
Deriving a Loss Function

Likelihood

$$p(y|w) = p\left(y \circ (X^\top w + \epsilon) > 0\right)$$

$$= \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, y \circ X^\top w,\, yy^\top \circ \Sigma\right) d\epsilon$$

$$= \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon$$



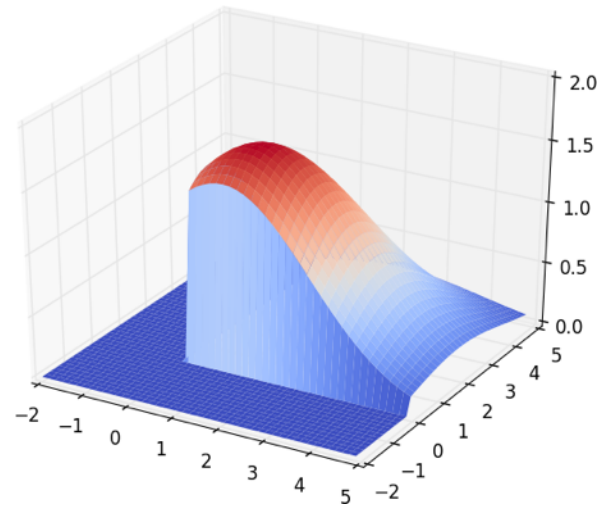
Deriving a Loss Function

Likelihood

$$p(y|w) = p\left(y \circ (X^\top w + \epsilon) > 0\right)$$

$$= \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, y \circ X^\top w,\, yy^\top \circ \Sigma\right) d\epsilon$$

$$= \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon$$

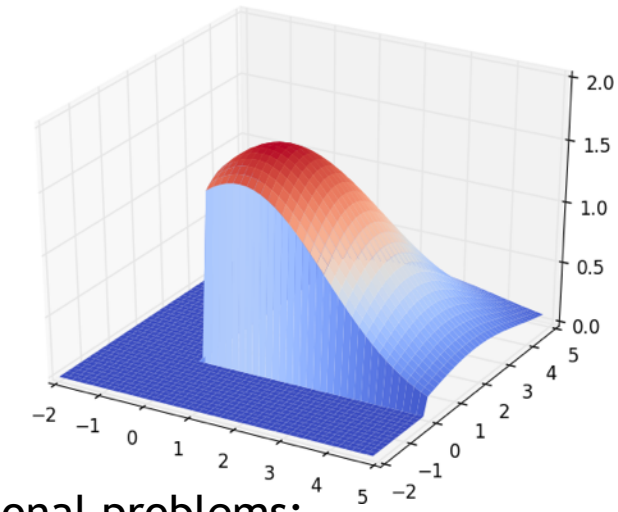


Loss Function (negative log posterior)

$$\mathcal{L}(w) \;=\; -\log \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon \;+\; \lambda_0 ||w||_1$$

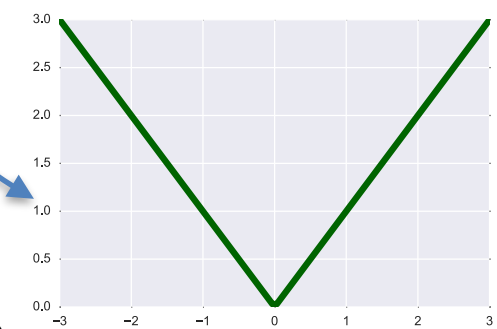
Minimizing the Loss Function

$$\mathcal{L}(w) \;=\; -\log \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon \;+\; \lambda_0 ||w||_1$$



Minimizing the objective function leads to two computational problems:

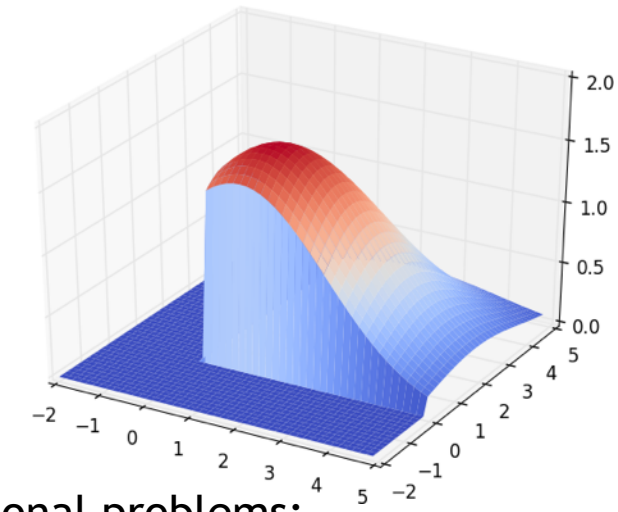
(i) intractable high-dimensional integral



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

Minimizing the Loss Function

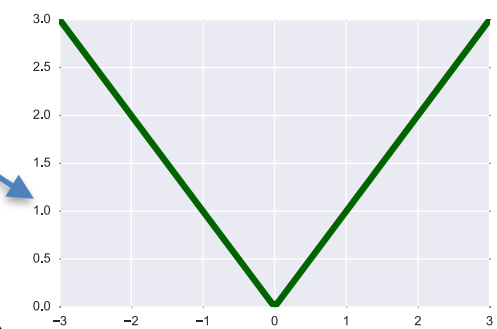
$$\mathcal{L}(w) \;=\; -\log \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon \;+\; \lambda_0 ||w||_1$$



Minimizing the objective function leads to two computational problems:

(i) intractable high-dimensional integral

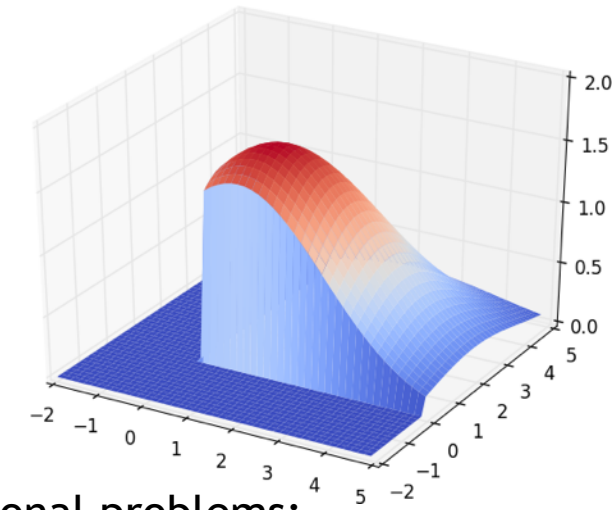
Solution: Expectation Propagation (EP)



(ii) l_1-norm regularizer is not everywhere differentiable

Minimizing the Loss Function

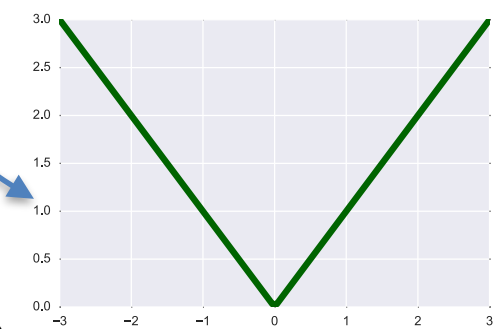
$$\mathcal{L}(w) \;=\; -\log \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon \;+\; \lambda_0 ||w||_1$$



Minimizing the objective function leads to two computational problems:

(i) intractable high-dimensional integral

⇒ **Solution:** **Expectation Propagation (EP)**

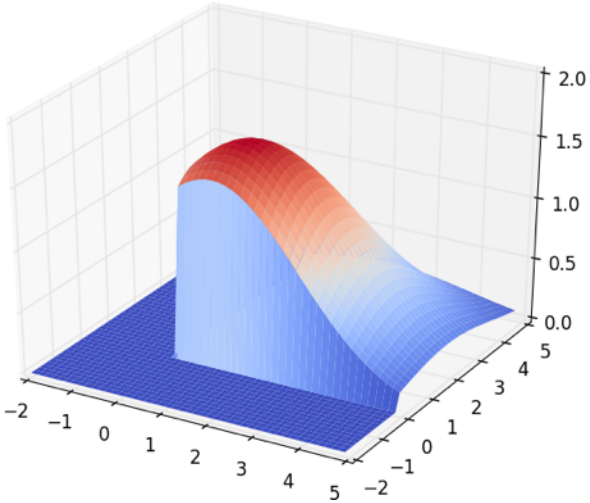


(ii) l_1-norm regularizer is not everywhere differentiable

⇒ **Solution:** **Alternating Direction Method of Multipliers (ADMM)**

The Likelihood Term

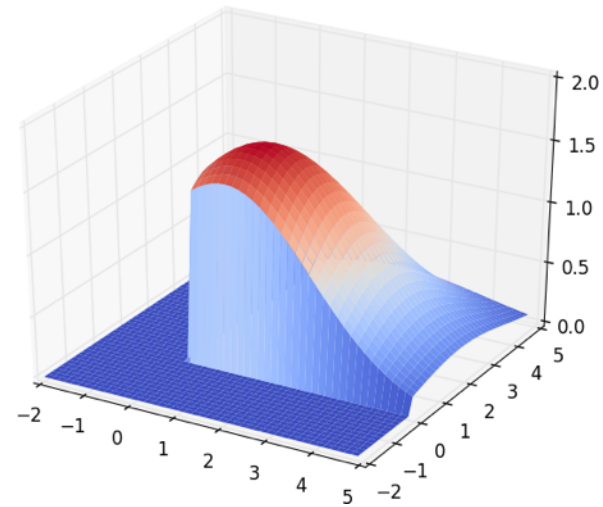
$$\mathcal{L}(w) \;=\; \underbrace{\boxed{-\log \int_{\mathbb{R}_+^n} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon}}_{\mathcal{L}_{\text{lik}}(w)} \;+\; \lambda_0 ||w||_1$$



How to optimize $\mathcal{L}_{\text{lik}}(w)$?

The Likelihood Term

$$\mathcal{L}(w) \;=\; \underbrace{\boxed{-\log \int_{\mathbb{R}_+^n} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon}}_{\mathcal{L}_{\text{lik}}(w)} \;+\; \lambda_0 ||w||_1$$

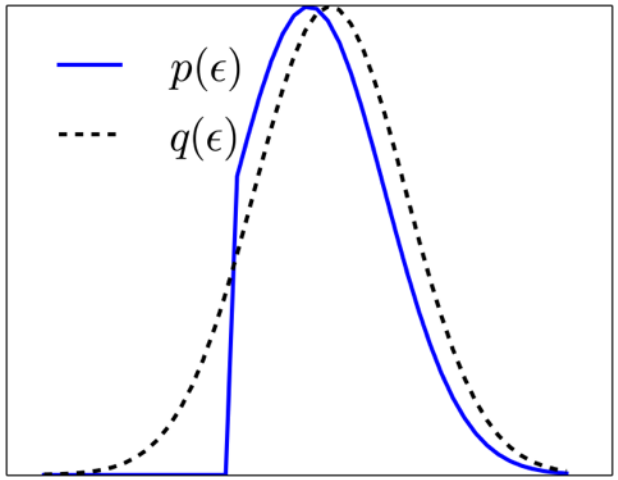


How to optimize $\mathcal{L}_{\text{lik}}(w)$?

- Want to apply second order minimization algorithm.
- Gradient and Hessian of $\mathcal{L}_{\text{lik}}(w)$ can be expressed in terms of the **1st and 2nd moment of the truncated Gaussian.**
- Problem: Computing the moments is still intractable.

Expectation Propagation

$$\mathcal{L}(w) = \underbrace{\boxed{-\log \int_{\mathbb{R}_+^n} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon}}_{\mathcal{L}_{\text{lik}}(w)} + \lambda_0 ||w||_1$$



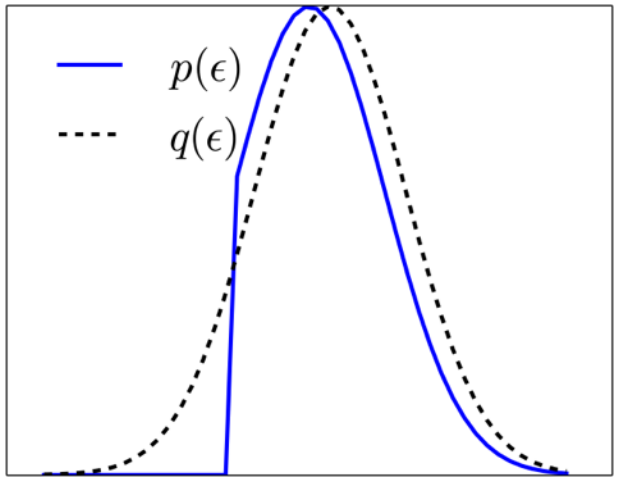
[J. Cunningham et. al., Gaussian probabilities and EP, arxiv 2011.]

Goal: Compute moments of (unnormalized) truncated Gaussian:

$$p(\epsilon; \mu, \tilde{\Sigma}) = \mathbf{1}_{\{\epsilon \in \mathbb{R}_+^n\}} \mathcal{N}(\epsilon; \mu, \tilde{\Sigma})$$

Expectation Propagation

$$\mathcal{L}(w) = \underbrace{\boxed{-\log \int_{\mathbb{R}_+^n} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon}}_{\mathcal{L}_{\text{lik}}(w)} + \lambda_0 ||w||_1$$



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

Goal: Compute moments of (unnormalized) truncated Gaussian:

$$p(\epsilon; \mu, \tilde{\Sigma}) = \mathbf{1}_{\{\epsilon \in \mathbb{R}_+^n\}} \mathcal{N}(\epsilon; \mu, \tilde{\Sigma})$$

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

Then: 1st and 2nd moment of *p* and *q* are approximately the same!

Optimizing the Likelihood Term

$$\mathcal{L}(w) \;=\; \underbrace{\boxed{-\log \int_{\mathbb{R}_+^n} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon}}_{\mathcal{L}_{\text{lik}}(w)} \;+\; \lambda_0 ||w||_1$$

Optimizing the likelihood term solved.

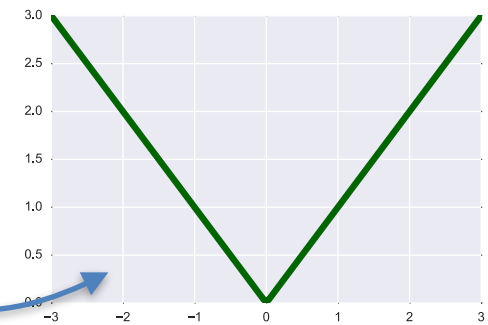
Optimizing the Likelihood Term

$$\mathcal{L}(w) \;=\; \underbrace{\boxed{-\log \int_{\mathbb{R}_+^n} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon}}_{\mathcal{L}_{\text{lik}}(w)} \;+\; \lambda_0 ||w||_1$$

Optimizing the likelihood term solved. ✓

Optimizing the Whole Objective

$$\mathcal{L}(w) \;=\; \boxed{-\log \int_{\mathbb{R}^n_+} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon \;+\; \lambda_0 ||w||_1}$$



ADMM

- Overcomes the problems of the non-differentiability of the regularizer.
- Essentially alternates between optimization updates corresponding to the likelihood term and the regularizer term.

Optimizing the Whole Objective

$$\mathcal{L}(w) \;=\; \boxed{-\log \int_{\mathbb{R}_+^n} \mathcal{N}\left(\epsilon;\, \mu(w),\, \tilde{\Sigma}\right) d\epsilon \;+\; \lambda_0 ||w||_1}$$

Inference Algorithm:

In each step:

- **Approximate** truncated Gaussian by an un-truncated Gaussian **via EP.**
- Use the 1st and 2nd moment of the approximation to compute the **gradient and Hessian of the Likelihood** term.
- Do ADMM **optimization step** corresponding to the **likelihood.**
- Do ADMM **optimization step** corresponding to the **regularizer.**

Summary

Correlated **Probit model:**

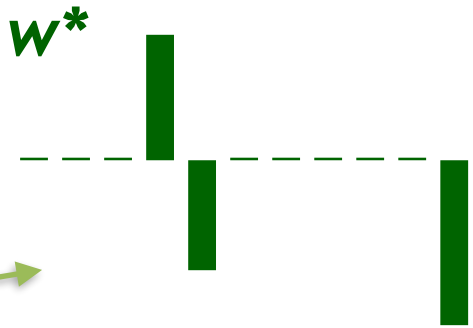
$$y_i = \mathrm{sign}(X_i^\top w + \epsilon_i) \qquad \epsilon \sim \mathcal{N}(0, \Sigma)$$

$$w \sim \mathrm{Laplace}(0, \lambda_0^{-1})$$

$$\Sigma = \lambda_1 \mathbf{I} + \lambda_2 X^\top X + \lambda_3 \Sigma_{\mathrm{side}}$$

Training:

$$w^* = \arg\max_w \, p(w|X, y)$$

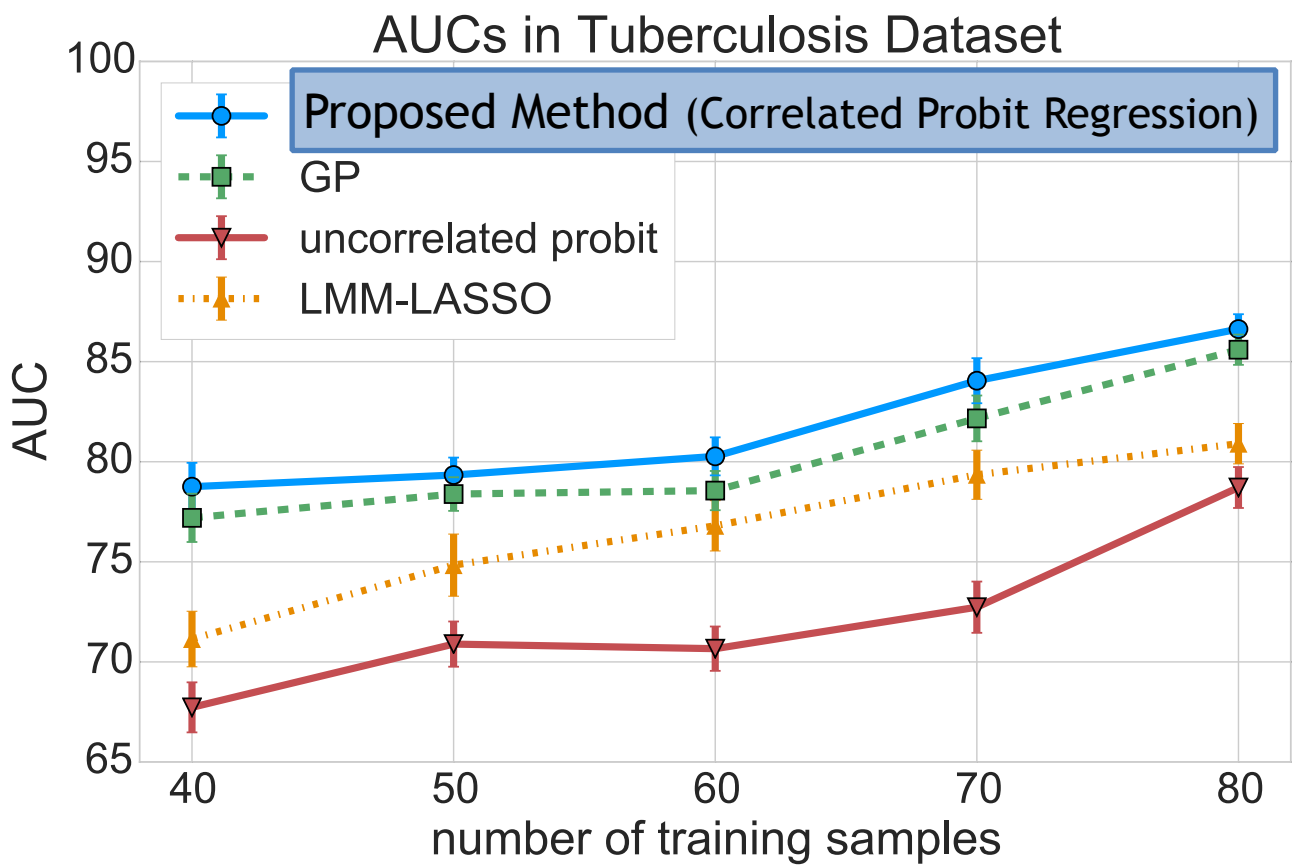


Selects genes which could be associated with the phenotype

Experiments (TBC)

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

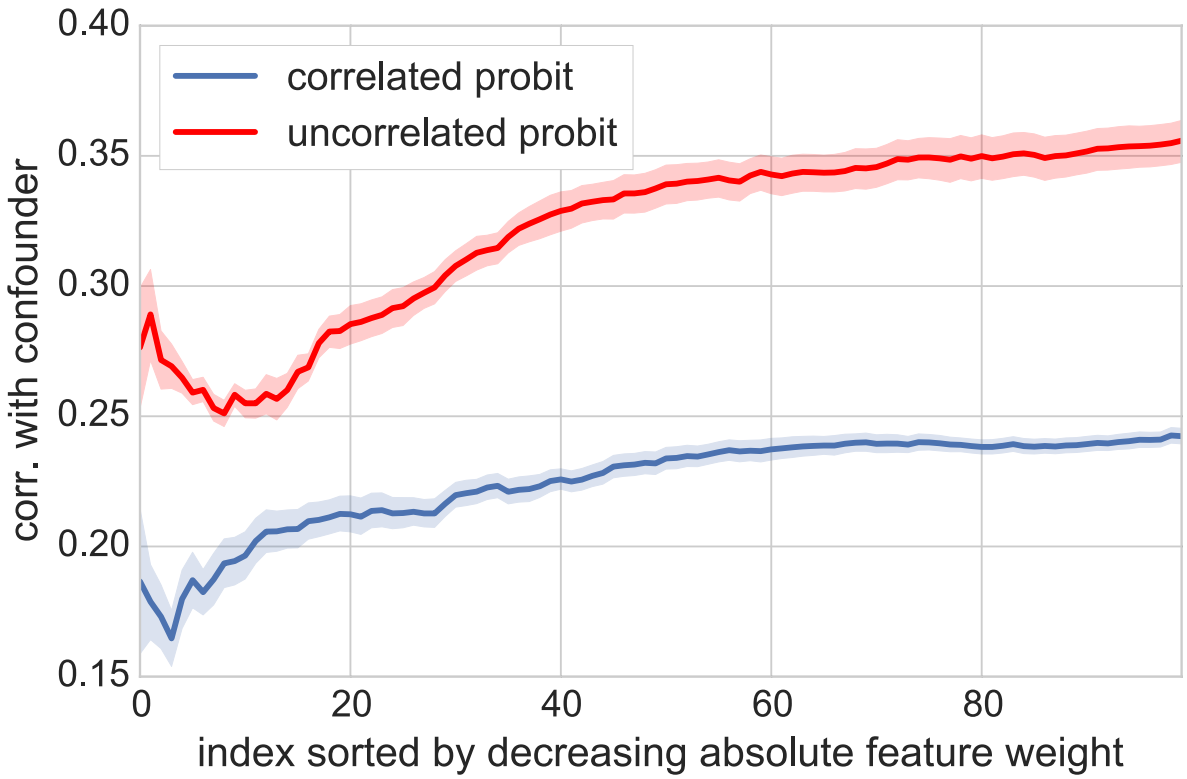
Tuberculosis data set:
Berry et. al., Nature 466, 2010.



Experiments (TBC)

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

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



More Experiments (in the paper)

- Good results on a **malware dataset**.

  (Here we correct for malware structure = similarities within a malware family)

- Experiments with **simulated data**.

Conclusion

Correlated Probit Regression

- **Novel algorithm for sparse feature selection** in binary classification, where the data are confounded, e.g. by population structure.
- We showed that the signals found by our model are **less correlated with the confounders.**
- Our method leads to **improved prediction performances** and lets us find sparse effects.
- Our method **scales up to high dimensions** (~500,000 features), but is only applicable to rather small datasets (~500 subjects).

Current Ongoing Research:

Scale it up!

From hundreds of subjects…
…to thousands of subjects.

Ongoing Research: Scalable Inference

Outlook:

- We use recent developments in scalable MCMC sampling: **Stochastic Gradient Fisher Scoring**.
- This method is based on using only mini batches of the data for each optimization step.
- Makes it applicable to larger datasets (**~10,000 subjects**).

Correlated Probit Regression

Paper (under review MLJ):

Preprint: "Sparse Estimation in a Correlated Probit Model."

arXiv:1507.04777

More Information and Papers:

www.florian-wenzel.de