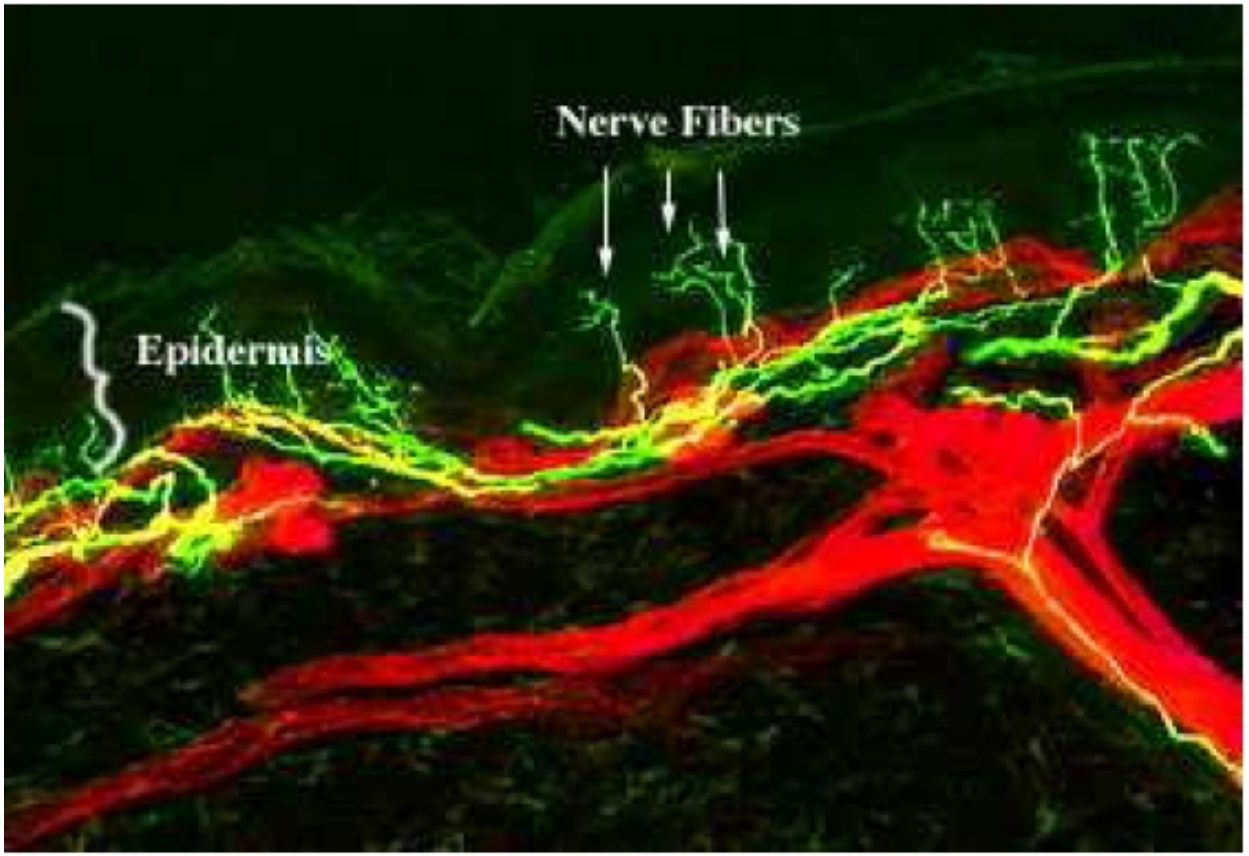
Spatial point processes: repetitions and non-spatial covariates

Epidermal nerve fibers

ENFs are thin nerve fibers in the epidermis (the outmost living layer of the skin)



Problem

Kennedy *et al.* (1999): Nerve fiber loss due to neuropathy does not seem to result in random removal of nerve trunks, rather the remaining nerves seem arranged in clusters
$\rightarrow$ Can we quantify this observation?

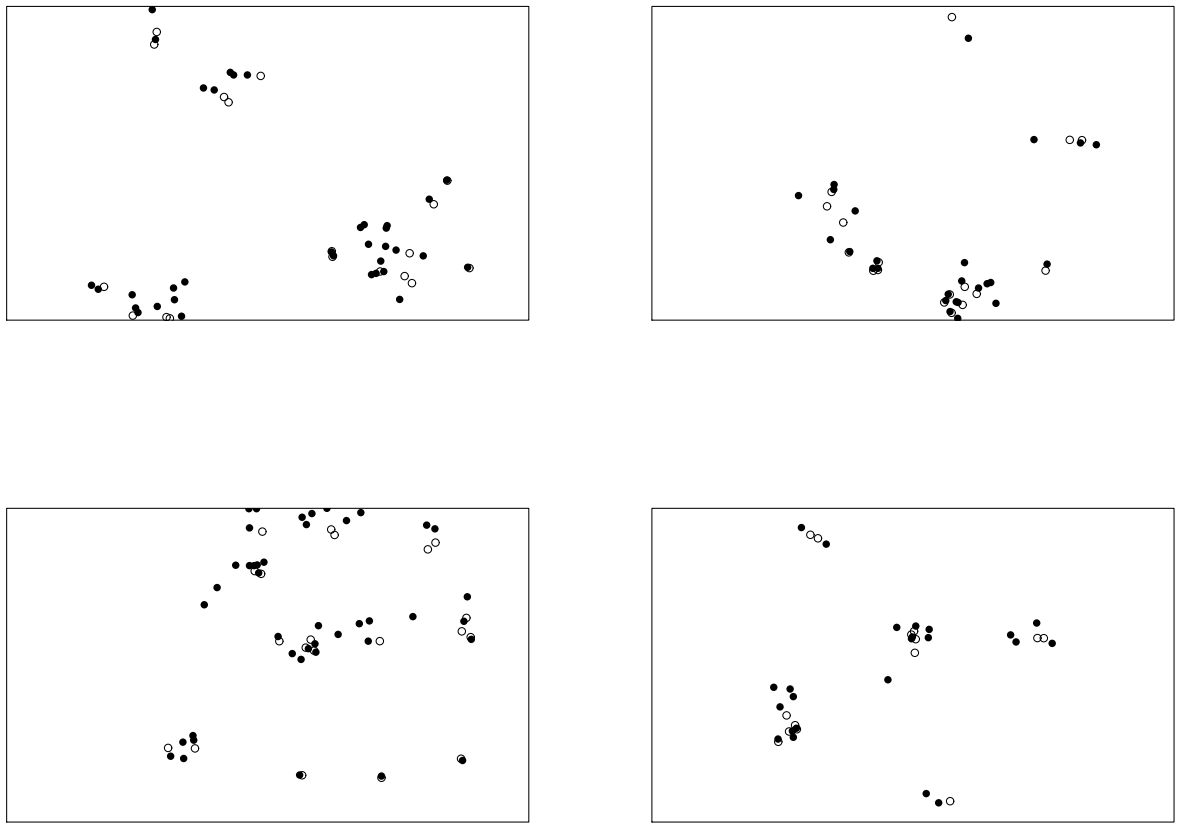
Is the spatial structure of ENFs affected by some (non-spatial) covariates, especially whether the subject is suffering from diabetic neuropathy?

How to include non-spatial covariates in the spatial analysis?

Data

- 32 healthy subjects and 15 subjects with (mild or moderate) diabetic neuropathy
- Two skin blisters (3-6 samples) from calf of each subject (replicates)
- Age, gender and body mass index (BMI) of each subject available

Spatial pattern of ENF entry (base) and end points



Locations of base points (open circles) and end points (small black dots) for one healthy subject (171)

Remark

The data are fibre patterns in 3D (with z direction much smaller than x and y directions) but we have looked only at the spatial pattern of (base points and) end points in 2D

- end points of ENFs sense heat and pain, and play, therefore, a more important role than the ENFs themselves (fibers can be omitted)
- our focus is on the spatial pattern of ENF coverage across the skin (2D projection appropriate)
- point patterns of ENF base and end points regarded as realizations of stationary spatial point processes

Our data are point patterns with replicates and non-spatial covariates

Second-order summary statistic

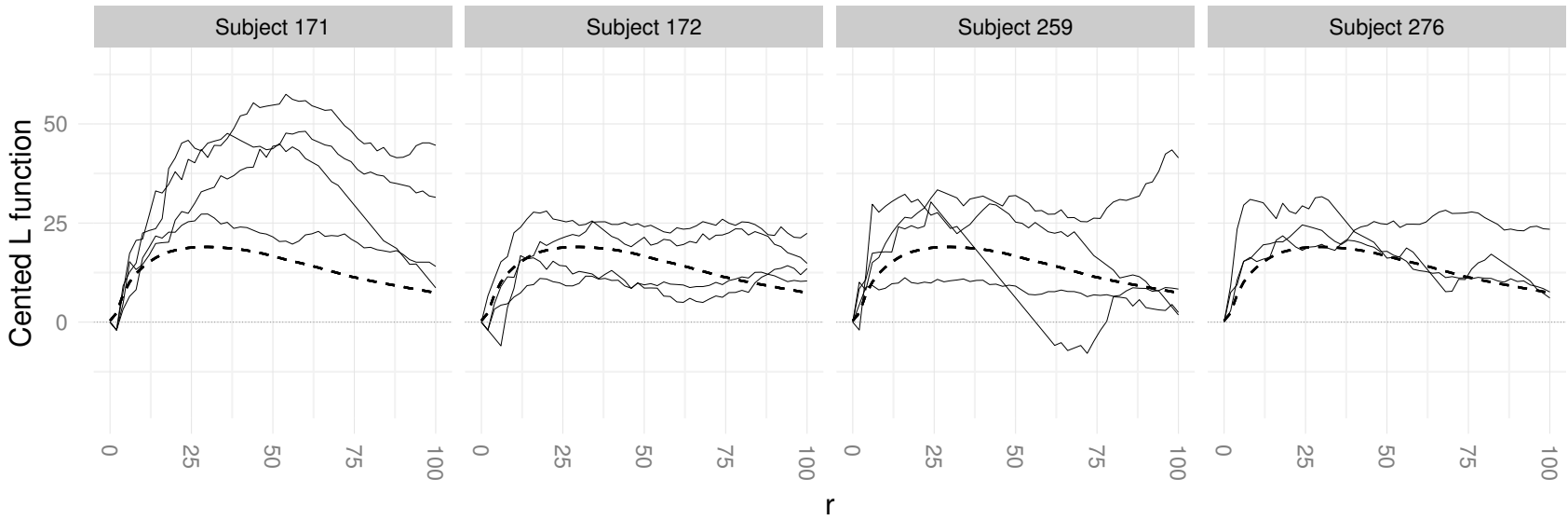
Ripley's K function: $\lambda K(r)$ is the expected number of further points within distance r from a typical point of the process, where λ is the intensity (mean number of points per unit area) of the process

We use a variance stabilizing and centered version of the K function, namely

$$L(r) - r = \sqrt{K(r)/\pi} - r,$$

which equals 0 under complete spatial randomness. Values less than zero indicate regularity and values larger than zero clustering.

Individual $L(r) - r$ functions for end points



Subject 171 and Subject 172 are healthy, the other two diseased

How to include non-spatial covariates?

- Pooled summary statistics for groups (Waller *et al.*, 2011)
- Summary statistic modeled by using linear mixed models (Myllymäki *et al.* 2012)
- Summary statistic modeled by using hierarchical Gaussian process regression (Myllymäki *et al.* 2013)

Linear mixed models for centered L function

L function modeled by using linear mixed models usually used to model growth curves, where

- distance r is the "time variable"
- fixed effects: disease status, age, gender, BMI, r, interactions between the covariates, interaction between the covariates and distance r
- random effects: intercept and r (both subject specific and sample specific)

Limitations

- We assumed linear dependence between the covariates and the characteristic (centered L function)
- $L(r) - r$ function modelled as a fourth order polynomial (somewhat ad hoc)

Hierarchical Gaussian process regression model for centered L function

- Flexible non-parametric models for making inference about the relationship between some characteristics (centered L function) and covariates
- We do not need to assume linear or any other particular form of dependence between the characteristics and covariates, a priori
- Bayesian approach

Hierarchical model

Our model is

$$y_{sjk} = L_{sj}(r_k) - r_k = f_1(\mathbf{x}_s, r_k) + f_2(s, r_k) + f_3(s, j, r_k) + \epsilon_{sjk},$$

where

- f_1 models the effect of age, gender, BMI and disease status (collected in $\mathbf{x}_s$) together with distance r
- f_2 models the subject-specific effect
- f_3 models the sample-specific effect
- latent function $f = f_1 + f_2 + f_3$
- ϵ_{sjk}'s are independent and $\sim N(0, \sigma^2)$

Hierarchical model: f_1 (covariates) and f_2 (subject-specific effect)

$$y_{sjk} = L_{sj}(r_k) - r_k = f_1(\mathbf{x}_s, r_k) + f_2(s, r_k) + f_3(s, j, r_k) + \epsilon_{sjk}$$

f_1 is a Gaussian process (GP) with

- mean $\overline{L(r_k) - r_k}$
- covariance function having an own length scale parameter for each covariate (age, gender, BMI, disease status) and for r
- values of f_1 are correlated within a subject but also between subjects due to similar covariate values.

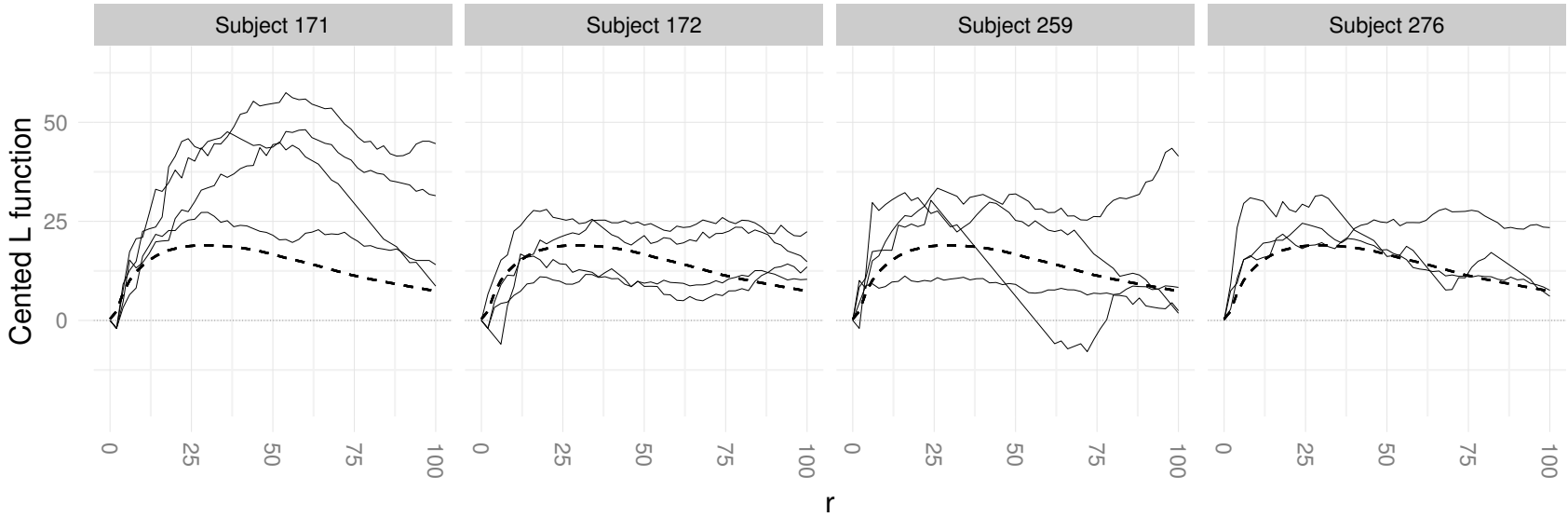
f_2 is a GP with

- mean zero
- covariance function, which is a priori the same for each subject

Hierarchical model: f_3 (sample-specific effect)

f_3 is a GP with

- mean zero
- covariance function, where the variance parameter σ^2_{3s} is allowed to vary from subject to subject
- values of f_3 are correlated only within a sample



Hierarchical model

- Observation model $\mathbf{y}|f, \sigma^2 \sim \prod_i N(y_i|f, \sigma^2)$
- GP prior

$$f(\mathbf{x})|\theta \sim GP(m(\mathbf{x}), k_1(\mathbf{x}, \mathbf{x}'|\theta_1) + k_2(\mathbf{x}, \mathbf{x}'|\theta_2) + k_3(\mathbf{x}, \mathbf{x}'|\theta_3))$$

- Hyperpriors

$$\sigma^2 \sim p(\sigma^2)$$
$$\theta_1 = (\phi_1, \sigma_1^2) \sim p(\phi_1)p(\sigma_1^2)$$
$$\theta_2 = (\phi_2, \sigma_2^2) \sim p(\phi_2)p(\sigma_2^2)$$
$$\theta_3 = \{\phi_3, \sigma_{3s}^2, s = 1, \ldots, N\} \sim p(\phi_3) \prod_{s=1}^{N} p(\sigma_{3s}^2|s_\sigma^2)$$

- Hyper-hyperprior $s_\sigma^2 \sim p(s_\sigma^2)$

Choices

- $r = 0, 12, 24, ..., 96$ (end points)
- Piecewice polynomial compactly supported covariance functions (less smooth for f_3 than for the first two components)
- Half-Student t and scaled inverse χ^2 priors for hyperparameters and for s_σ^2

Bayesian inference and posterior predictive L functions

Since f and the likelihood are Gaussian, we can integrate out the latent function and obtain the *log marginal likelihood*

$$\log p(\mathbf{y}|\mathbf{X}, \theta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|K + \sigma^2 I| - \frac{1}{2}\mathbf{y}^T(K + \sigma^2 I)^{-1}\mathbf{y},$$

where $\theta = (\theta_1, \theta_2, \theta_3)$ collects all the parameters of f and K is the covariance matrix.

The posterior distribution of the latent function f_1

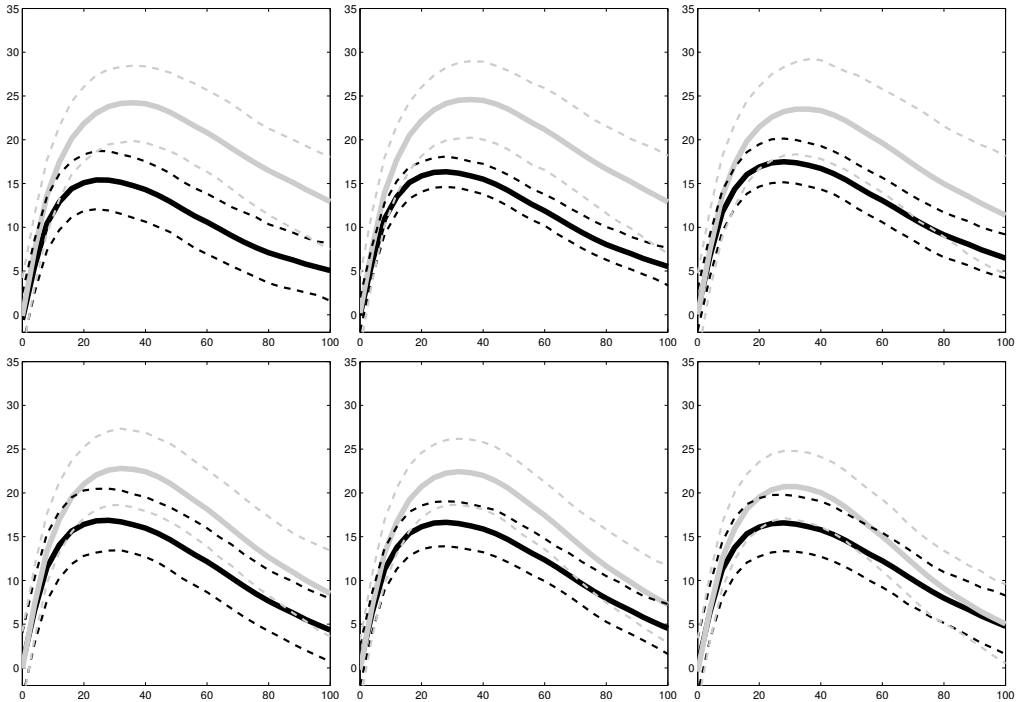
$$p(f_1|\mathbf{y}, \mathbf{X}) = \int p(f_1|\mathbf{y}, \mathbf{X}, \theta, \sigma^2)p(\theta, \sigma^2|\mathbf{y}, \mathbf{X})\mathrm{d}\theta\mathrm{d}\sigma^2$$

can be obtained by Monte Carlo integration over the hyperparameters

Computations

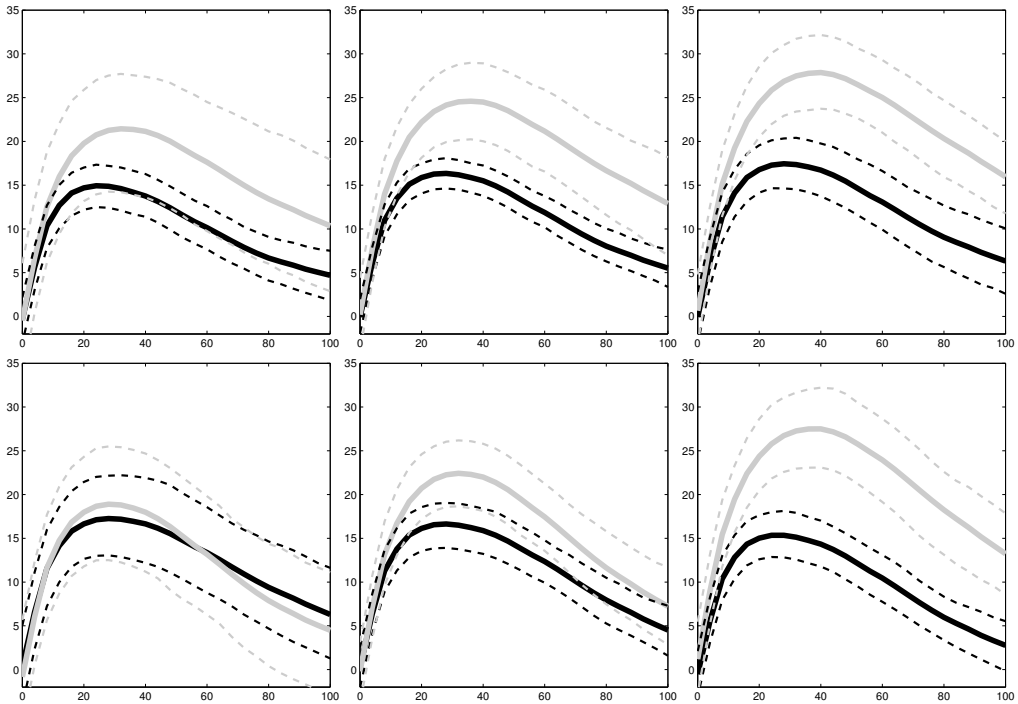
- To obtain the posterior distribution of the parameters, we run an MCMC simulation updating in turns the hyper-hyperparameter and the hyperparameters
- For sampling the Matlab toolbox GPstuff (Vanhatalo et al., 2013) is used

Mean prediction centred L curves (mean of the posterior predictive distribution of f_1) for end points



Female (first row), male (second row)
From left to right: Age 30, 45, 60; BMI is fixed to 25
Healthy (black), diseased (grey)

Mean prediction centred L curves for end points



Female (first row), male (second row)
From left to right: BMI 20, 25, 30; Age is fixed to 45
Healthy (black), diseased (grey)

Results

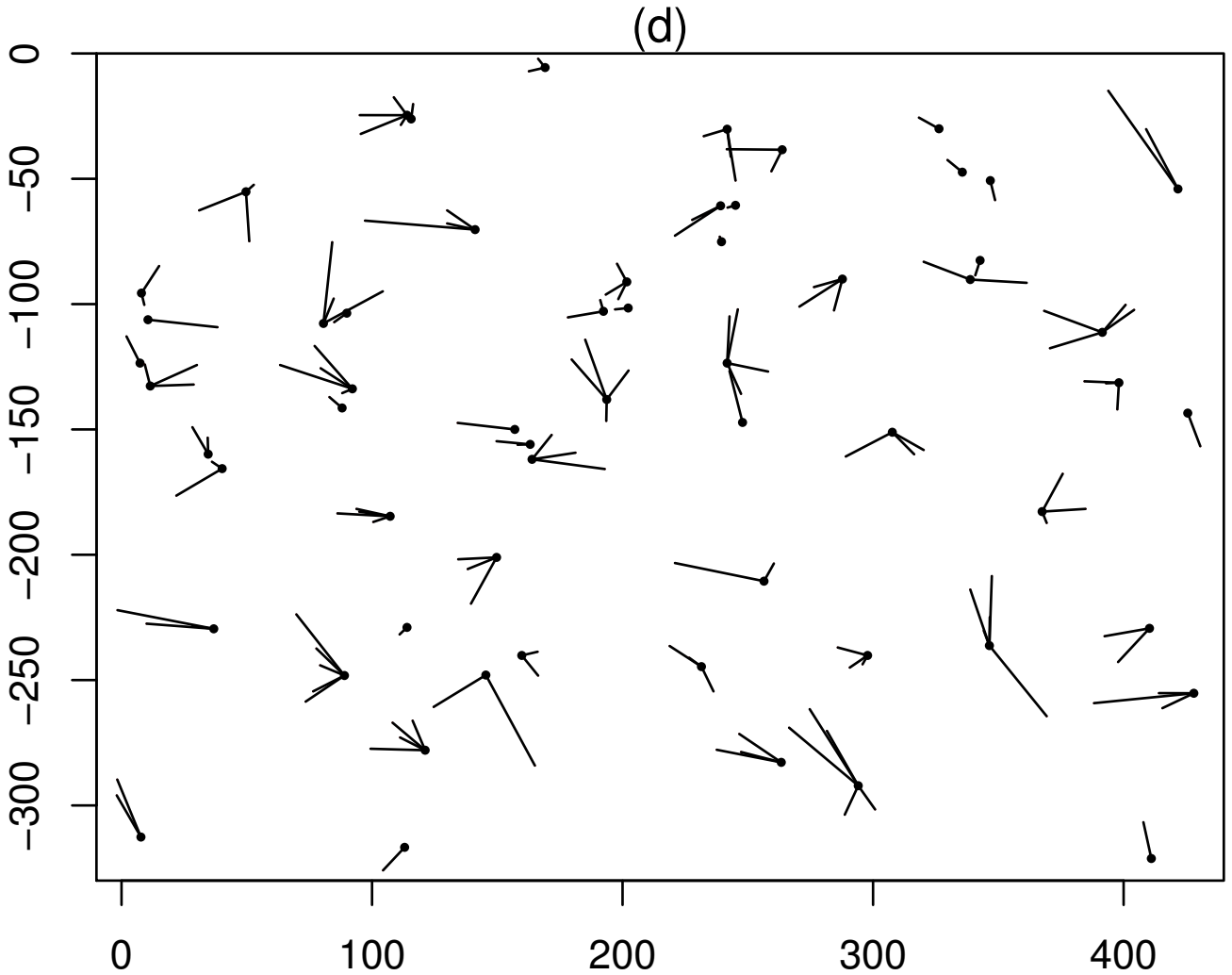
Base points: covariates (including disease status) do not seem to have any effect on the ENF pattern

End points

- diseased patterns clearly more clustered than healthy ones
- difference between healthy and diseased patterns is more clear for women than for men
- difference between healthy and diseased patterns is more easily seen for younger subjects and subjects with high BMI than for older subjects and subjects with low BMI
- effects of age, gender and BMI not evident

Preliminary 2D point process models

We know which end points are connected to which base



Preliminary 2D model

Non-orphan cluster (NOC) model:

- Base points are assumed to be a realization of a Poisson process (data from thigh)
- End points
  - number of end points: (e.g.) Poisson distribution
  - distance from each end point to its entry point: Gamma distribution
  - direction of the end point: von Mises distribution, where angles opposite to the nearest neighbouring entry point are favoured

Reference: Olsbo *et al.* (2013)

Some remarks

- Model the locations of the base points first, and given a realization of base points, the end points are modeled (hierarchical structure)
- Locations of end points may not be independent of each other as in the preliminary model

Some things to think about

- Compare different body parts (thigh, calf, foot)
- 3D spatial point (and fibre) process models to incorporate further details in ENF structure
- Spatio-temporal models for ENF growth
- How to use replicates?
- How to include non-spatial covariates in the models?

References

- Kennedy, W.R., Nolano, M., Wendelschafer-Crabb, G., Johnson, T.L., Tamura, E. A skin blister method to study epidermal nerves in peripheral nerve disease, Muscle and Nerve 22 (1999) 360-371.
- Myllymäki, M., Panoutsopoulou, I.G., Särkkä, A. Analysis of spatial structure of epidermal nerve entry point patterns based on replicated data, Journal of Microscopy 247 (2012) 228-239.
- Myllymäki, M., Särkkä, A., Vehtari, A. Hierarchical second-order analysis of replicated spatial point patterns with non-spatial covariates. To appear in *Spatial Statistics*.
- Olsbo, V., Myllymäki, M., Waller, L.A. and Särkkä, A. Development and evaluation of spatial point process models for epidermal nerve fibers. *Mathematical Biosciences* 243 (2013) 178-189.

References (continues)

- Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., Vehtari, A. Gpstuff: Bayesian modeling with Gaussian processes, Journal of Machine Learning Research 14 (2013) 1175-1179.
- Waller, L.A., Saärkkaä, A., Olsbo, V., Myllymaäki, M., Panoutsopoulou, I.G., Kennedy, W.R., Wendelschafer-Crabb, G. Second- order spatial analysis of epidermal nerve fibers. *Statistics in Medicine* 30 (2011) 2827-2841.