Pre-Symptomatic Prediction of Plant Drought Stress Using Dirichlet-Aggregation Regression on Hyperspectral Images

Kristian Kersting[†•*] **Zhao Xu**[†•] **Mirwaes Wahabzada**[†] **Christian Bauckhage**[†] **Christian Thurau**[‡]
[†]Fraunhofer IAIS, Sankt Augustin, Germany [‡]Game Analytics Aps., Copenhagen, Denmark

Christoph Roemer[◦] **Agim Ballvora**[§] **Uwe Rascher**[⋆] **Jens Leon**[§] **Lutz Pluemer**[◦]
[⋆]Inst. of Bio- and Geosciences,IBG-2: Plant Sciences, Forschungszentrum Juelich, Germany
[§] Inst. of Crop Science and Resource Conservation, Plant Breeding; [◦]Inst. of Geodesy and Geoinformation
University of Bonn, Germany

Abstract

Pre-symptomatic drought stress prediction is of great relevance in precision plant protection, ultimately helping to meet the challenge of "How to feed a hungry world?". Unfortunately, it also presents unique computational problems in scale and interpretability: it is a temporal, large-scale prediction task, e.g., when monitoring plants over time using hyperspectral imaging, and features are 'things' with a 'biological' meaning and interpretation and not just mathematical abstractions computable for any data. In this paper we propose Dirichlet-aggregation regression (DAR) to meet the challenge. DAR represents all data by means of convex combinations of only few extreme ones computable in linear time and easy to interpret. Then, it puts a Gaussian process prior on the Dirichlet distributions induced on the simplex spanned by the extremes. The prior can be a function of any observed meta feature such as time, location, type of fertilization, and plant species. We evaluated DAR on two hyperspectral image series of plants over time with about 2 (resp. 5.8) Billion matrix entries. The results demonstrate that DAR can be learned efficiently and predicts stress well before it becomes visible to the human eye.

Introduction

Water scarcity is a principle global problem that causes aridity and serious crop losses in agriculture. It has been estimated that drought can cause a depreciation of crop yield up to 70% in conjunction with other abiotic stresses (Boyer 1982; Pinnisi 2008). Climate changes and a growing human population in parallel thus call for a sincere attention to advance research on understanding of plant adaptation under drought. A deep knowledge of the adaptation process is essential in improving management practices, breeding strategies as well as engineering viable crops for a sustainable agriculture in the coming decades. Accordingly, there is a dire need for crop cultivars with high yield and strong resistance against biotic and abiotic stresses.

Unfortunately, understanding stress is not an easy task. Stress resistance is the result of a complex web of interactions between the genotype and the environment leading to phenotypic expressions. It is contributed by a number of related traits that are controlled mostly by polygenic inheritance. In the past, a slow progress in the development of improving cultivars was mainly due to poor understanding of genetic factors that impact tolerance to drought (Passioura 2002). Recently, progress has been made in understanding the genetic basis of drought related quantitative trait loci (QTL), see e.g. (Lebreton et al. 1995; McKay et al. 2008). More recently, OMICS approaches have offered a direct molecular insight into drought tolerance mechanism, see e.g. (Rabbani et al. 2010; Guo et al. 2010; Abdeen, Schnell, and Miki 2010). However, genetic and biochemical approaches are time consuming and still fail to fully predict the performance of new lines in the field. In recent years it is discussed that phenomic approaches, that measure the structural and functional status of plants may overcome the limited predictability and some authors have attributed this lack of high throughput phenomic data as the "phenomic bottleneck" (Richards et al. 2010).

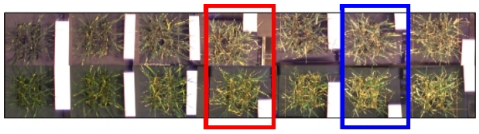

Figure 1: Which of the plants (one per row) suffers from drought? Can we predict it from the first few hyperspectral images only? Shown are the hyper-spectral images (640x640x69) projected to RGB (640x640x3). Time progresses from left to right. The blue box denotes when the symptoms typically become visible to the human eye. As indicated by the red box, DAR can predict drought stress earlier; about 1.5 weeks earlier. (Best viewed in color)

Hyper-spectral imaging provides a particularly promising approach to sensor-based phenotyping. Its measurements were observed to contain early indicators of plant stress, see e.g. (Rascher et al. 2007; Rascher and Pieruschka 2008). In contrast to conventional cameras, which record only 3 wavelengths per pixel, hyper-spectral cameras record a spectrum of several hundred wavelengths ranging from approximately 300nm to 2500nm resulting in big data cubes. These spectra contain information as to changes of the pigment composition of leaves which are the result of

[*]Contact author: kristian.kersting@iais.fraunhofer.de. • Both authors contributed equally.

metabolic processes involved in plant responses to biotic or abiotic stresses. This information can be used e.g. using SVMs for classification of hyper-spectral signatures and in turn for prediction of biotic stress before symptoms become visible to the human eye, see e.g. (Rumpf et al. 2010; Römer et al. 2010). These and similar studies, however, are considerably different from the goal of the present paper: multi-step ahead prediction of drought in time based on sequences of hyper-spectral images as also illustrated in Fig. 1. Here, recorded image spectra, i.e., pixels in the hyper-spectral images are not annotated, as even stressed plants show only local signs of stress. We evaluate all measured wavelength of many hyper-spectral images and not only a few for one or two images. This poses a challenge in scale since the amount of phenotyping data produced, easily grows into TeraBytes if several plants are monitored over time. For instance, our datasets have in total about 2 (resp. 5.8) Billion matrix entries. Finally, since phenotyping is necessarily interdisciplinary, requiring that scientists with complementary skills work together, it is desirable to obtain results and models, which can intuitively be interpreted by researchers who are not machine learning and data mining experts.

This makes it difficult to use off-the-shelf statistical techniques such as PCA, HMMs, SVMs and GPs directly on sequences of hyper-spectral images: many of them assumed labeled input data; they typically do not scale well with the amount of data if no form of approximation is used that is often accompanied by information loss; they often do not provide easy-to-interpret features/models; they make assumptions on the true generating distribution, which we do not know; and they assume data points as input. Our input objects, however, are best described by multiple samples gathered in data matrix, i.e., by "data clouds", see e.g. (Davis and Dhillon 2006).

Consequently, we propose a novel prediction approach, called Dirichlet-aggregation regression (DAR). It does not make any assumption on the generating distribution of each data matrix. Instead, DAR employs a recent linear time, data-driven matrix factorization approach to represent the data clouds by means of convex combinations of only few extreme data samples across all clouds and time steps. This new representation imposes a natural distribution on the data, namely the distribution on the simplex spanned by the extreme data samples. This was recently proven to be successful (Kersting et al. 2012) for detecting drought stress patterns and contrasts to standard approaches e.g. using local features evaluated at certain keypoints of hyperspectral images, see e.g. (Mukherjee, Velez-Reyes, and Roysam 2009), where the true generating distribution is not known, or just assuming Gaussians, see e.g. (Davis and Dhillon 2006), which is likely to be not true in our application at hand. Moreover, in contrast to histogram-based "data clouds" approaches, see e.g. (Sakurai et al. 2008), one can perform naturally Bayesian inference to quantify the "impact" of extremes on a dispersion model over time. This is exactly the main technical contribution of the present paper[1].

Specifically, DAR puts a Gaussian process prior on the Dirichlet distributions induced on the simplex spanned by the extremes. The prior can be a function of any arbitrary types of observed continuous, discrete and categorical features such as time, location, fertilization, and plant species with no additional coding, yet inference remains relatively simple. As our experimental results show, by just using time as meta feature, DAR can already predict the drought stress of plants well and before it become visible to the human eye. Prediction models of this kind have great potential as they provide better insights into early stress reactions and to identify the most relevant moment when biologists have to gather samples for invasive, molecular examinations.

We proceed as follows. We start off by briefly reviewing how to use matrix factorization to estimate distributions over phenotypes and to detect drought patterns. Then, we introduce DAR and show how to use it for predicting drought multiple steps ahead in time. Before concluding, we present our experimental results.

Dirichlet Aggregation of Phenotypes

We briefly recall some fundamentals of interpretable matrix factorization and how they lead to parametric probability distributions over phenotypes and to a formal notion of drought level. For more details, please refer to (Goreinov and Tyrtyshnikov 2001; Frieze, Kannan, and Vempala 2004; Mahoney and Drineas 2009) and the recent application to drought detection (Kersting et al. 2012).

Scientists working on plant phenotyping regularly confront the problem of finding meaningful patterns hidden in massive, high-dimensional, and temporal observations. Consider e.g. our experiments. Hyper-spectral images of resolution 640x640x69 were taken of 10 (resp. 12) plants l at 7 (resp. 20) different days t. Each image can be viewed as a data matrix $\boldsymbol{X}^{t,l} \in \mathbb{R}^{m \times n}$ with m =640x640 and $n = 69$, i.e., with about 28 Million entries[2]. Horizontally stacking all data matrices per experiment results in a single matrix with about 2 (resp. 5.8) Billion matrix entries. So, how can we find easy-to-interpret patterns for drought level prediction in these sequences of data matrices?

One natural candidate are matrix factorization techniques that factorize $\boldsymbol{X}$ into two matrices $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$ where the matrix of basis elements $\boldsymbol{W} \in \mathbb{R}^{m \times k}$ and the coefficient matrix $\boldsymbol{H} \in \mathbb{R}^{k \times n}$ are typically determined from minimizing the squared Frobenius norm $\left\|\mathbf{X} - \mathbf{WH}\right\|^2$. They allow one to embed high dimensional data $\boldsymbol{X}$ in lower dimensional spaces $\boldsymbol{H}$ and can therefore mitigate effects due to noise, uncover latent relations, or facilitate further processing and ultimately help finding patterns in the data set distribution. One prominent approach e.g. consists in truncating the Singular Value Decomposition (SVD), which expresses the data in terms of linear combinations of the top singular vectors. While these basis vectors are optimal in a statistical sense, the SVD has been criticized for it is less faithful to the nature

[1] A similar approach is known for learning topic models (Mimno and McCallum 2008; Wahabzada, Xu, and Kersting 2010). In the present paper, we do not employ topic models and additionally consider the multi-step prediction problem.

[2] For sake of readability, we will often omit the indices t, l.

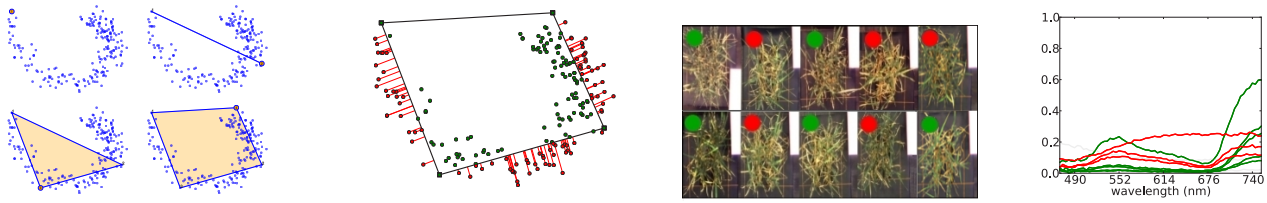

Figure 2: Fast plant phenotyping using SiVM. From left to right. **(1)** 2D example illustrating how SiVM determines four extreme points. **(2)** Any sample point can be expressed as a convex combination of these extremes. While points inside of the simplex can be reconstructed exactly, points on the outside are approximated by their projection onto its closest facet. **(3)** Images of all plants in year 2010 at the fourth measurement day. **(4)** Examples of extreme spectra found by SiVM. They can be grouped into "dry" (red) and "healthy" (green), see main text for details. A single spectrum essentially shows how much light at each wavelength is reflected from the plant spot/pixel it corresponds to. (Best viewed in color)

of the data at hand. For instance, the data mining practitioner — as in our application — often tends to assign a "physical" meaning to the resulting factors. Such reification must be based on an intimate knowledge of the application domain and can often not be justified from mathematics. This also holds for other techniques such as NMF and kMeans. More importantly, classical approaches confront us with the difficulty of characterizing sophisticated patterns of data point distributions in a unified parametric and interpretable form. This is generally intractable (Wang, Zha, and Qin 2007).

An alternative are interpretable factorization approaches. Here, the basis vectors $\boldsymbol{W}$ are c columns selected from $\boldsymbol{X}$ that maximize their volume $Vol(\mathbf{W}^{m\times c}) = |\det \mathbf{W}|$, and $\boldsymbol{H}$ is restricted to convexity, i.e., $h_{ij} \geq 0$ and $\sum_i h_{ij} = 1$. Consequently, the basis vectors have naturally a biological meaning since they have been observed and can easily be interpreted by a domain expert. However, the maximum-volume criterion is provably NP-hard (Civril and Magdon-Ismail 2009). An approximation, called Simplex Volume Maximization (SiVM), was recently introduced by Thurau *et al.* (2012) and empirically proven to be quite successful. For a subset $\mathbf{W}$ of c columns from $\mathbf{X}$, let $\Delta(\mathbf{W})$ denote the $c-1$-dimensional simplex formed by the columns in $\mathbf{W}$. Now, the volume of the c-simplex $\mathrm{Vol}(\Delta(\mathbf{W}))$ is $\mathrm{Vol}(\Delta(\mathbf{W}))_c^2 = \theta \det \boldsymbol{A}$ where $\theta = \frac{-1^{c+1}}{2^c(c!)^2}$ and $\det \mathbf{A}$ is the so-called *Cayley-Menger* determinant (Blumenthal 1953). It is computed from a matrix of distances between points using an $\mathcal{O}(c \cdot n)$ efficient greedy algorithm. Fig. 2 illustrates SiVM and shows extreme signatures found when running it on hyperspectral images of plants.

However, how does SiVM help us devising probability distributions over the phenotypes? From a geometric point of view, the columns $\mathbf{h}_1, \ldots, \mathbf{h}_n$ of $\boldsymbol{H}$ can be considered as data points residing in a simplex spanned by the selected extreme signatures $\mathbf{W}$ so that there are natural parametric distributions for $\mathbf{h}_i$ on the simplex. Probably the best known one is the Dirichlet: $\mathcal{D}(\mathbf{h}_i|\boldsymbol{\alpha}) = B(\boldsymbol{\alpha}) \prod_{j=1}^c h_{ij}^{\alpha_j - 1}$ where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_c)$. The normalization constant has value $B(\boldsymbol{\alpha}) = \Gamma(S(\boldsymbol{\alpha}))/\prod_{j=1}^c \Gamma(\alpha_j)$, where Γ denotes the gamma function and $S(\boldsymbol{\alpha}) = \sum_{j=1}^c \alpha_j$. Dirichlets, naturally enforce our convexity constraint on the reconstructions, $0 \leq h_{ij} \leq 1$ and $\sum_{j=1}^c h_{ij} = 1$, so that changing one h_{ij} impacts all other h_{ik}. To estimate the parameters $\boldsymbol{\alpha}$ from the reconstructions one can follow a maximum-likelihood approach. The exact details are not important for the present paper and we refer to (Minka 2000). More important is that the distributional view on the hyper-spectral data provides us with an intuitive measure for drought stress: the expected probability of observing a healthy spot, which we call the "drought stress level" of a plant.

More formally, given the $\boldsymbol{\alpha}$ estimated from the hyper-spectral images, we note that the marginal distribution of the j-th reconstruction dimension follows a Beta distribution $\mathcal{D}(\alpha_j, S(\boldsymbol{\alpha}) - \alpha_j)$ so that the expected value of the j-th reconstruction dimension is $\mu_j = E[\alpha_j] = \alpha_j / S(\boldsymbol{\alpha})$. Intuitively, this means that each α_j controls "aggregation" of mass of reconstructions near the corresponding selected column c_j, which also explain the term "Dirichlet aggregation" regression. Now assume that we have labeled each dimension as either "background", "healthy" or "dry". Averaging the expected values for each "healthy" resp. "dry" dimension and treating them as parameters of a Beta distribution yields the drought stress level of a plant.

But how do we classify signatures into "background", "healthy" or "dry"? This follows from the results in (Kersting et al. 2012). A signature is "background" if the corresponding pixel is not a leaf spot. This is easy to verify using the original images since all extremes are observed hyper-spectral pixels. If it is not background, it is "healthy" if it has a low reflectance (< 0.1) in the wavelengths 470nm - 540nm (chlorophyll a+b): it is a green spot. If this is not the case and the ratio of maximal reflectances observed at the wavelengths 470nm - 540nm (i.e., they are or turn brown) and at the wavelengths 700nm and above is > 0.5 (they start to overheat): it is "dry". See also Fig. 2**(4)**.

However, recall that we are actually interested in the levels over time. Again following (Kersting et al. 2012), we first select extreme columns on the horizontal stack of all data matrices. This captures global dependencies as we represent the complete data by means of convex combinations extreme data points selected across all time steps. Then, on the simplex spanned by the extreme points, one estimates Dirichlet distributions specified by $\boldsymbol{\alpha}^{t,l}$ over all reconstruc-

tions per day t and plant l. This captures local dependencies. Finally, we compute the drought levels using the $\boldsymbol{\alpha}^{t,l}$.

In the following, we show how to extend this detection approach to the significantly more important prediction task.

Bayesian Dirichlet-Aggregation Regression

Reconsider a particular reconstruction $\boldsymbol{h}^{t,l}$ computed by SiVM for single pixel in hyperspectral image of a plant l at time t. We take a Bayesian perspective and assume that $\boldsymbol{h}^{t,l}$ was generated from a hidden Dirichlet distribution (denoted as $\mathcal{D}$) parameterized by the variable $\alpha^{t,l} = [\alpha_1^{t,l}, \ldots, \alpha_c^{t,l}]^T$. Assuming a Gaussian process prior on the $\boldsymbol{\alpha}$s for a single plant, the joint probability of the reconstructions and the hidden Dirichlet aggregation[3] $\boldsymbol{\eta}^t = \log \boldsymbol{\alpha}^t$ at time $t = 1, \ldots, \tau$ can be written as $P(\{h^t, \eta^t\}_{t=1}^{\tau} | \mathbf{K}) =$

$$\prod_{c=1}^{C} \mathcal{N}(\eta_c^{1:\tau} | \mathbf{K}) \prod_{t=1}^{\tau} \prod_{j=1}^{M} \mathcal{D}(\boldsymbol{h}_j^t | \alpha_{1:C}^t) \; . \qquad (1)$$

That is, the likelihood of the reconstructions is a Dirichlet distribution with $\boldsymbol{\alpha}_{1:C}^t = [\exp(\eta_1^t), \ldots, \exp(\eta_C^t)]^T$, and $\eta_c^{1:\tau} = [\eta_c^1, \ldots, \eta_c^\tau]^T$ is drawn from a Gaussian distribution (denoted as $\mathcal{N}$) with covariance matrix $\mathbf{K}$. The covariance function generating $\mathbf{K}$ can be any Mercer kernel function. In our experiments, we used the well known squared exponential (SE) $k(x_i, x_j) = \kappa^2 \exp(-\frac{\rho^2}{2} \sum_s^S (x_{i,s} - x_{j,s})^2)$ with two hyperparameters $\vartheta = (\kappa, \rho)$.

The major inference problem we are facing now is to estimate the hidden Dirichlet aggregation $\boldsymbol{\eta}^t$ for each plant given the prior parameters K and the observations $\boldsymbol{h}^t$. The basic idea to achieve this is to find the $\hat{\boldsymbol{\eta}}^{1:\tau}$ that maximizes the logarithm of the complete data likelihood $P(\{\boldsymbol{h}^t, \boldsymbol{\eta}^t\}_{t=1}^{\tau} | \mathbf{K})$ in Eq. (1), i.e., $\hat{\eta}^{1:\tau} = \arg\max_{\{\boldsymbol{\eta}^1, \ldots, \boldsymbol{\eta}^\tau\}} \log P(\{h^t, \boldsymbol{\eta}^t\}_{t=1}^{\tau} | \mathbf{K})$. The log-likelihood can be written as $\mathcal{L} = \log P(\{\boldsymbol{h}^t, \boldsymbol{\eta}^t\}_{t=1}^{\tau} | \mathbf{K}) =$

$$\begin{aligned} &\sum_{c=1}^{C} -\frac{\tau}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{K}| - \frac{1}{2} (\eta_c^{1:\tau})^T \mathbf{K}^{-1} (\boldsymbol{\eta}_c^{1:\tau}) \\ &+ \sum_{t=1}^{\tau} \sum_{j=1}^{M} \Big[\log \Gamma(\sum_{c=1}^{C} \exp(\eta_c^t)) - \sum_{c=1}^{C} \log \Gamma(\exp(\eta_c^t)) \\ &+ \sum_{c=1}^{C} (\exp(\eta_c^t) - 1) \log h_{j,c}^t \Big] \qquad (2) \end{aligned}$$

where $\Gamma(\cdot)$ denotes the Gamma function. Note that Eq. (2) is the log-likelihood of a single plant. The plant-specific Dirichlet aggregation $\eta^{1:\tau}$ sequenses are independent of each other given the common Gaussian process prior.

We train the model using a coordinate descent on the η and the hyper-parameters of the Gaussian process prior. The partial derivative of $\mathcal{L}$ w.r.t. $\boldsymbol{\eta}$ can be written as $\frac{\partial \mathcal{L}}{\partial \eta_c^t} =$

$$-\sum_{t'=1}^{\tau} \eta_c^{t'} K_{t,t'}^{-1} + \alpha_c^t \sum_{j=1}^{M} \left[\log h_{j,c}^t + \psi(\sum_{c'=1}^{C} \alpha_{c'}^t) - \psi(\alpha_c^t) \right]$$

with $\alpha_c^t = \exp(\eta_c^t)$, j indexing the SiVM reconstructions of a hyper-spectral image, and $\psi(\cdot)$ denotes the digamma function. To compute the partial derivative $\mathcal{L}$ w.r.t. ϑ we note

[3]Taking the logarithm ensures that the estimated αs are positive.

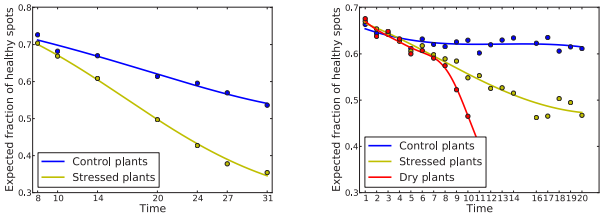

Figure 3: Dirichlet-aggregation regression of drought levels over time in year 2010 **(left)** and in year 2011 **(right)** using all hyperspectral images available. Colors indicate controlled/stressed plants. (Best viewed in color)

that the Gaussian process prior is shared by all the plants, thus optimizing the hyper-parameters ϑ accounts for the log-likelihood of all plants. So actually, we can consider to find $\hat{\vartheta} = \arg\max_{\vartheta} \sum_{l=1}^{L} \log P(\{h^{t,l}, \boldsymbol{\eta}^{t,l}\}_{t,l} | \vartheta)$. The partial derivative now simplifies to $\frac{\partial \mathcal{L}}{\partial \vartheta_i} =$

$$\frac{1}{2} \sum_{l=1}^{L} \sum_{c=1}^{C} (\boldsymbol{\eta}_c^{l,1:\tau})^T \mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \vartheta_i} \mathbf{K}^{-1} (\eta_c^{l,1:\tau}) - \frac{LC}{2} Tr(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \vartheta_i}),$$

where $Tr(\cdot)$ is the trace of a matrix and the matrix derivative $\partial \mathbf{K} / \partial \vartheta_i$ can be computed as $\frac{\partial K_{t,t'}}{\partial \vartheta_i} = \frac{\partial}{\partial \vartheta_i} k(t, t'; \vartheta)$. For example, with the SE kernel used in our experiments, $\partial K_{t,t'} / \partial \vartheta_i$ is $2\kappa \exp(-\frac{\rho^2}{2} D_{t,t'}^2)$ for $\vartheta_i \equiv \kappa$; and $-\rho \kappa^2 D_{t,t'}^2 \exp(-\frac{\rho^2}{2} D_{t,t'}^2)$ for $\vartheta_i \equiv \rho$. Here, $D_{t,t'}$ denotes the distance in time between t and t'.

To summarize, Dirichlet-aggregation regression (DAR) performs the following two steps until convergence:

1. Optimize the logarithm of the complete likelihood w.r.t. the hidden Dirichlet aggregations η_c^t for each plant.
2. Optimize the log-likelihood of all plants w.r.t. the hyper-parameters ϑ of the common Gaussian process prior.

Fig. 3 shows the drought levels estimated by DAR averaged over each group of plants on the two datasets used in our experiments. As one can see, DAR nicely smoothes SiVMs "hard" drought level (shown as dots).

Our main objective, however, is to predict drought stress ahead in time and not to detect it only.

Bayesian Prediction of Drought Stress Levels

Having a Bayesian regression model at hand, one can employ an iterative method for prediction consisting in making repeated one-step ahead predictions, up to the desired horizon. For the one-step ahead prediction at time t^*, we simply fall back on the standard equations for Gaussian process regression, see e.g. (Rasmussen and Williams 2006). That is, the expectation of the hidden Dirichlet aggregation η_c^{t*} of a plant is computed as $\hat{\eta}_c^{t*} = \mathbf{K}_* \mathbf{K}^{-1} \hat{\eta}_c^{1:\tau}$, where $\mathbf{K}_*$ denotes the covariances between the new time t_* and the known ones $t = 1, \ldots, \tau$ computed with the kernel function using the estimated hyper-parameters $\hat{\vartheta}$. Given $\hat{\eta}_c^{t*}$, we compute the drought stress level as described in the previous section.

For the multiple-step ahead prediction task we follow the method proposed by Girard et al. (2002). That is, we predict the next time step ahead, using the estimate of the output of the current prediction, as well as previous outputs (up to some lag U), as the input to the prediction of the next time step, until the prediction k steps ahead is made. Thus, the prediction k steps ahead is a random vector with mean formed by the predicted means of the lagged outputs.

More formally, we assume a state-space model on the drought levels ξ^{t_k}, i.e., $\xi^{t_k} = g(x^{t_k}) + \epsilon$ with $x^{t_k} = [\xi^{t_k-1}, \ldots, \xi^{t_k-U}]$, where x^{t_k} is the state at time t_k composed of previous criteria (up to the Lag U), and ϵ is a Gaussian noise. Thus, the prediction at time t_k is a function $g(\cdot)$ of the current state, i.e. previous prediction corrupted by Gaussian noise. Now, we employ a Gaussian process to solve the system. As Girard et al. argue, the benefit of doing so are twofold: (1) There is no assumption on the mathematical form of the function $g(\cdot)$. It is modeled as an arbitrary function drawn from a Gaussian process. (2) The uncertainty induced by successive predictions is taken into account. This leads to more realistic predictive uncertainties by modeling the lagged prediction outputs of previous steps as noisy inputs to a Gaussian process.

More formally, the mean of ξ^{t_k} can we write as $\hat{\xi}^{t_k} = \mathbf{q}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$, where $\mathbf{y}^T = [\xi^1, \xi^2, \ldots, \xi^\tau]$ is composed of previous predictions. The covariance matrix $\boldsymbol{\Sigma}$ with $\Sigma_{ij} = c(\mathbf{x}^{t_i}, \mathbf{x}^{t_j})$ correlates the prediction outputs ξ^{t_i} and ξ^{t_j}. We again use the SE kernel. q_i is the covariance between ξ^{t_k} and ξ^{t_i} with the input uncertainty Σ^{t_k}

$$q_i = |\mathbf{W}^{-1}\Sigma^{t_k} + I|^{-1/2} \times \exp(-0.5(\hat{\mathbf{x}}^{t_k} - \mathbf{x}^{t_i})^T(\mathbf{W} + \Sigma^{t_k})^{-1}(\hat{\mathbf{x}}^{t_k} - \mathbf{x}^{t_i})),$$

where $\mathbf{W}$ is a $U \times U$ diagonal matrix with $W_{i,i} = 1/\rho^2$, which is the length scale parameter in the SE kernel. The term $\hat{\mathbf{x}}^{t_k}$ denotes the expected state at the time t_k, which consists of previous mean predictions. Finally, the variance of the prediction at the time t_k can be written as:

$$v^{t_k} = c(\hat{\mathbf{x}}^{t_k}, \hat{\mathbf{x}}^{t_k}) + Tr(\boldsymbol{\Sigma}^{-1}(\mathbf{y}\mathbf{y}^T\Sigma^{-1} - I)\mathbf{Q}) - Tr(\hat{\xi}^{t_k})^2$$

with $\log Q_{ij} =$

$$-\frac{1}{2}\log|2\mathbf{W}^{-1}\Sigma^{t_k} + I| - \frac{1}{4}(\mathbf{x}^{t_i} - \mathbf{x}^{t_j})^T\mathbf{W}^{-1}(\mathbf{x}^{t_i} - \mathbf{x}^{t_j}) - \frac{1}{2}(\hat{\mathbf{x}}^{t_{ij}} - \hat{\mathbf{x}}^{t_k})^T(\frac{1}{2}\mathbf{W} + \Sigma^{t_k})^{-1}(\hat{\mathbf{x}}^{t_{ij}} - \hat{\mathbf{x}}^{t_k}),$$

where $\hat{\mathbf{x}}^{t_{ij}} = (\mathbf{x}^{t_i} + \mathbf{x}^{t_j})/2$.

Now, we have everything together to predict drought stress levels of plants from hyper-spectral images: **(A)** Using SiVM, we compute few extreme signatures, say 50, in the temporal fashion explained above and classify them accordingly **(B)** On the simplex spanned by the extremes, we estimate the latent Dirichlet aggregation values per plant and time step using DAR. **(C)** Using the Gaussian process over the latent Dirichlet aggregation values, we compute the drought levels of each plant and time step using the classification of extreme spectra "background", "healthy" and "dry". **(D)** Finally, we predict drought levels multiple steps ahead in time using the just described Gaussian process approach.

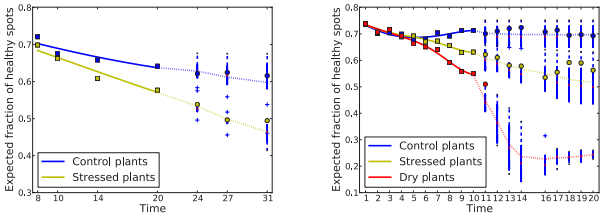

Figure 4: Bayesian drought level prediction. **(Left)** Predictions (levels over time) for year 2010 and **(Right)** year 2011. In both experiments, the drought levels of the second half of measurement days were predicted based on a DAR model trained (including the extraction of extreme spectra) on the data gathered in the first half of measurement days. Colors indicate controlled/stressed plants. (Best viewed in color)

Rainout Shelter Experiments

Our intention here is to investigate the following questions: **(Q1)** Can DAR predict drought stress pre-symptomatically from hyperspectral images? **(Q2)** Can DAR smoothing lead to improved detection of drought stress patterns compared to the ones using SiVM only?

To this aim, we implemented DAR in Python. Following (Kersting et al. 2012), we ran SiVM using a variant of the Kolmogorov-Smirnov distance, i.e., we treated the reflectance signatures as empirical distributions.

Datasets: We considered two sets of hyperspectral imaging data. Both datasets were recorded under semi-natural conditions in a rainout shelter. For the controlled water stress in the rainout shelter three barley summer cultivars Scarlett and Wiebke and Barke were chosen for the water stress in this study. The experiments were performed in rain-out shelters at the experimental station of our University. The seeds were sown in 11.5 liter pots filled with 17.5 kg of substrate Terrasoil (Cordel&Sohn, Salm, Germany). In year 2010, the first dataset, the genotype Scarlett was used in two treatments (well-watered and with reduced water) with 6 pots per treatment. In year 2011, the second dataset, the genotypes Wiebke and Barke were used in pot experiments arranged in a randomized complete block design with three treatments (well-watered and two drought stressed) with 4 pots per genotype and treatment. The drought stress was induced either by reducing the total amount of water or by the complete withholding of water. In both cases the stress was started at developmental stage BBCH31 (Biologische Bundesanstalt, Bundessortenamt and Chemical industry). By reducing the irrigation the water potential of substrate remained at the same level as in the well-watered plots for the first seven days but decreased rapidly in the following 10 days reaching 40% compared to the control. For the measurements the plants were transferred in the laboratory and illumination was provided by 6 halogen lamps (400 W ECO, OSRAM, Munich, Germany) fixed at distance of $1,6$ meter from the support where the pots where placed to take the pictures.). The camera was mounted at the same level as the lamps in NADIR position. In year 2010 images were taken at 10

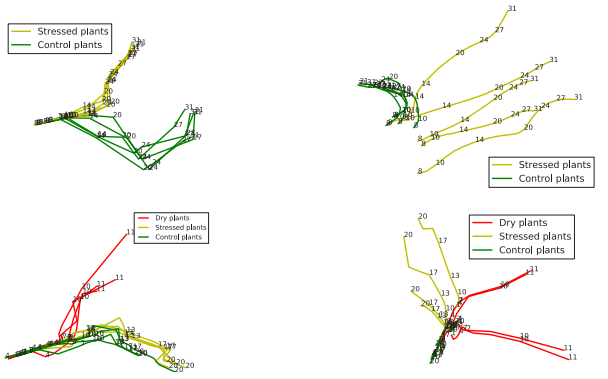

Figure 5: Improved drought stress detection using DAR. From left to right: **(1)** Dirichlet traces for year 2010 without and **(2)** with DAR smoothing. **(3)** Dirichlet traces for year 2011 without and **(4)** with DAR smoothing. Colors indicate controlled/stressed plants; numbers denote the measurement days. (Best viewed in color)

time-points, twice per week starting from day four of water-stress. This gave us 70 data cubes of resolution 640x640x69. We transform each cube into a dense 'pixel x spectra' matrix. This resulted in 70 data matrices of resolution 640x640x69. Stacking them horizontally together resulted in a dense data matrix with about 2 Billion entries. In year 2011 images were taken every consecutive day starting at the second day of watering reduction. Images were taken at 11 time-points for the non-irrigated plants and at 20 time-points for plants with reduced water amount. Following the same procedure as for year 2010, this resulted in about 5.8 Billion entries.

Learning Setup: To investigate **(Q1)**, we split the data of year 2011 (resp. 2010) into the first half, denoted as 2011.A (resp. 2010.A), and the second half, denoted as 2011.B (resp. 2010.B). Then, we extracted 50 extreme signatures from 2011.A (resp. 2010.A) and learned a DAR regression model on 2011.A (resp. 2010.A). The resulting hyper-parameters are $\vartheta = (1.73, 0.19)$ (resp. $(2.02, 0.01)$). Then, we classified the extreme signatures into "healthy", "dry", and "background" as described earlier, computed drought levels for 2011.A (resp. 2010.A) based on the estimated DAR model and used them to predict the drought levels for 2011.B (resp. 2010.B). To investigate **(Q2)**, we used the complete year 2010 (resp. 2011) data to learn a DAR model and computed Euclidean embeddings as described in (Kersting et al. 2012) using the smoothed $\boldsymbol{\alpha}$s.

Experimental Results: Fig. 4 summarizes the prediction results. In both cases, the notches of the boxes of the drought levels predicted for the different groups of plants at the last measurement day do not overlap. This offers evidence of a statistically significant difference between the predicted medians. Moreover, the predictions match the "SiVM-only" values well. We conclude that DAR can fully distinguish all groups of plants. Thus, as in the fully observed case, we can group the extreme spectra based on their probability in the different groups at the last measurement days. The extreme signatures found were essentially identical to the ones found in (Kersting et al. 2012) as already shown in Fig. 2. We conclude that the classified signatures indeed conform to plant physiologically knowledge. Moreover, since the symptoms do not become visible to the human eye at this time, question **(Q1)** can be answered affirmatively.

Fig. 5 summarizes the detection results. As one can see, the Euclidean embeddings based on the DAR smoothing are more reasonable. The healthy plants stay close together in a small region. The stressed plants stay close totgether only in the early days; in later days they diverge due to the dispersion of senescence. And, the differences in senescence development between the different groups of plants is pronounced. Thus, **(Q2)** can also be answered affirmatively.

Finally, we note that running SiVM can be parallelized so that the plant phenotyping runs in just about 30 minutes (Kersting et al. 2012). Estimating the DAR model and making the predictions is a matter of minutes.

Conclusions

Early stress prediction is of great relevance in precision plant protection. Pre-symptomatic water stress detection is of particular interest, ultimately helping to meet the challenge of "How to feed a hungry world?" (Editorial 2010). In this context, hyper-spectral image sensors are an established, sophisticated method for early stress detection. However, they gather massive, high dimensional data clouds over time, which together with the demand of physical meaning of the prediction model present unique computational problems in scale and interpretability. Motivated by this, we introduced Dirichlet-aggregation and presented the — to the best of our knowledge — first application of AI techniques to early drought stress prediction based on hyperspectral image sequences. Our experimental results on two large-scale plant phenotyping datasets demonstrate that the estimated temporal models are meaningful, conform to existing plant physiological knowledge, and are fast to compute. This is an encouraging sign that the vision of high throughput precision phenotyping is not insurmountable. Detailed measurements of plant characteristics can be analyzed at massive scale to collectively provide estimates of trait phenotypes for many of the underlying genotypes that comprise a typical plant breeding population.

Our work provides several interesting avenues for future work. Next to experiments under field conditions e.g. in an experimental agricultural site, one should aim at improving prediction quality even further by developing hierarchical, (semi-)supervised and relational versions of DAR. Active Bayesian regression approaches could provide better insights into early stress reactions and identify the most relevant moment when biologists have to gather samples for invasive, molecular examinations.

Acknowledgements: The authors would like to thank the reviewers for their feedback. The work was partially supported by the Fraunhofer ATTRACT fellowship STREAM and by the German Federal Ministry of Education and Research (BMBF) under the project number BMBF/315309/CROP.SENSe.

References

Abdeen, A.; Schnell, J.; and Miki, B. 2010. Transcriptome analysis reveals absence of unintended effects in drought-tolerant transgenic plants overexpressing the transcription factor abf3. *BMC Genomics* 11. Published online on January 28, doi: 10.1186/1471-2164-11-69.

Blumenthal, L. M. 1953. *Theory and Applications of Distance Geometry*. Oxford University Press.

Boyer, J. 1982. Plant productivity and environment. *Science* 218:443–448.

Civril, A., and Magdon-Ismail, M. 2009. On Selecting A Maximum Volume Sub-matrix of a Matrix and Related Problems. *TCS* 410(47–49):4801–4811.

Davis, J., and Dhillon, I. 2006. Differential entropic clustering of multivariate gaussians. In *Proceedings of Neural Information Processing Systems (NIPS)*, 337–344.

Editorial. 2010. How to Feed a Hungry World. *Nature* 7306(466):531–532.

Frieze, A.; Kannan, R.; and Vempala, S. 2004. Fast monte-carlo algorithms for finding lowrank approximations. *Journal of the ACM* 51(6):1025–1041.

Girard, A.; Rasmussen, C.; Quiñonero Candela, J.; and Murray-Smith, R. 2002. Gaussian process priors with uncertain inputs — application to multiple-step ahead time series forecasting. In *Proceedings of Neural Information Processing Systems (NIPS-02)*.

Goreinov, S., and Tyrtyshnikov, E. 2001. The maximum-volume concept in approximation by low-rank matrices. In *Contemporary Mathematics*, volume 280. 47–51.

Guo, P.; Baum, M.; Grando, S.; Ceccarelli, S.; Bai, G.; Li, R.; von Korff, M.; Varshney, R.; Graner, A.; and Valkoun, J. 2010. Differentially expressed genes between drought-tolerant and drought-sensitive barley genotypes in response to drought stress during the reproductive stage. *Journal of Experimental Botanic* 60:3531–3544.

Kersting, K.; Wahabzada, M.; Roemer, C.; Thurau, C.; Ballvora, A.; Rascher, U.; Leon, J.; Bauckhage, C.; and Pluemer, L. 2012. Simplex distributions for embedding data matrices over time. In *Proceedings of the 12th SIAM International Conference on Data Mining (SDM)*.

Lebreton, C.; Lazic-Jancic, V.; Steed, A.; Pekic, S.; and Quarrie, S. 1995. Identification of qtl for drought responses in maize and their use in testing causal relationships between traits. *Journal of Experimental Botanic* 46:853–865.

Mahoney, M., and Drineas, P. 2009. CUR matrix decompositions for improved data analysis. *PNAS* 106(3):697–702.

McKay, J.; Richards, J.; Sen, S.; Mitchell-Olds, T.; Boles, S.; Stahl, E.; Wayne, T.; and Juenger, T. 2008. Genetics of drought adaptation in arabidopsis thaliana ii. qtl analysis of a new mapping population , kas-1 x tsu-1. *Evolution* 62:3014–3026.

Mimno, D., and McCallum, A. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Minka, T. 2000. Estimating a Dirichlet distribution. In *A note publically available from the author's homepage*.

Mukherjee, A.; Velez-Reyes, M.; and Roysam, B. 2009. Interest points for hyperspectral image data. *IEEE T. Geoscience and Remote Sensing* 47(3):748–760.

Passioura, J. 2002. Environmental biology and crop improvement. *Functional Plant Biology* 29:537–554.

Pinnisi, E. 2008. The blue revolution, drop by drop, gene by gene. *Science* 320:171–173.

Rabbani, M.; Abe, K. M. H.; Khan, M.; Katsura, K.; Ito, Y.; Yoshiwara, K.; Seki, M.; Shinozaki, K.; and Yamaguchi-Shinozaki, K. 2010. Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cdna microarray and rna gel-blot analyses. *Plant Physiology* 133:1755–1767.

Rascher, U., and Pieruschka, R. 2008. Spatio-temporal variations of photosynthesis: The potential of optical remote sensing to better understand and scale light use efficiency and stresses of plant ecosystems. *Precision Agriculture* 9:355–366.

Rascher, U.; Nichol, C.; Small, C.; and Hendricks, L. 2007. Monitoring spatio-temporal dynamics of photosynthesis with a portable hyperspectral imaging system. *Photogrammetric Engineering and Remote Sensing* 73:45–56.

Rasmussen, C. E., and Williams, C. K. I. 2006. *Gaussian Processes for Machine Learning*. The MIT Press.

Richards, R.; Rebetzke, G.; Watt, M.; Condon, A.; Spielmeyer, W.; and Dolferus, R. 2010. Breeding for improved water productivity in temperate cereals: phenotyping, quantitative trait loci, markers and the selection environment. *Functional Plant Biology* 37(2):85–97.

Römer, C.; Bürling, K.; Rumpf, T.; Hunsche, M.; Noga, G.; and Plümer, L. 2010. Robust fitting of fluorescence sprectra for presymptomatic wheat leaf rust detection with Support Vector Machines. *Computers and Electronics in Agriculture* 79(1):180–188.

Rumpf, T.; Mahlein, A.-K.; Steiner, U.; Oerke, E.-C.; and Plümer, L. 2010. Early Detection and Classification of Plant Diseases with Support Vector Machines Based on Hyperspectral Reflectance. *Computers and Electronics in Agriculture* 74(1):91–99.

Sakurai, Y.; Chong, R.; Li, L.; and Faloutsos, C. 2008. Efficient distribution mining and classification. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 632–643.

Thurau, C.; Kersting, K.; Wahabzada, M.; and Bauckhage, C. 2012. Descriptive matrix factorization for sustainability: Adopting the principle of opposites. *Journal of Data Mining and Knowledge Discovery* 24(2):325—354.

Wahabzada, M.; Xu, Z.; and Kersting, K. 2010. Topic models conditioned on relations. In *Proceedings of ECML PKDD–10*.

Wang, H.-Y.; Zha, H.; and Qin, H. 2007. Dirichlet Aggregation. In *Proc. of ICML-2007*.